BIASED AND UNBIASED CROSS-VALIDATION IN DENSITY ESTIMATION

David W. Scott and George R. Terrell

Technical Report No. 23

APRIL 1986

Laboratory for Computational Statistics

Department of Statistics
Stanford University

Biased and Unbiased Cross-Validation in Density Estimation

by

David W. Scott and George R. Terrell

Rice University

and

Stanford University

Abstract

Nonparametric density estimation requires the specification of smoothing parameters. The demands of statistical objectivity make it highly desirable to base the choice on properties of the data set. In this paper we introduce some biased cross-validation criteria for selection of smoothing parameters for kernel and histogram density estimators. These criteria are obtained by estimating L_2-norms of derivatives of the unknown density and provide slightly biased estimates of the average squared-L_2 error or mean integrated squared error. These criteria are roughly the analog of the generalized cross-validation procedure for orthogonal series density estimators. We present the relationship of the biased cross-validation procedure to the least squares cross-validation procedure, which provides unbiased estimates of the mean integrated squared error. Both methods are shown to be based on U-statistics. We compare the two methods by theoretical calculation of the noise in the cross-validation functions and corresponding cross-validated smoothing parameters, by Monte Carlo simulation, and by example. Surprisingly large gains in asymptotic efficiency are observed between biased and unbiased cross-validation when the underlying density is sufficiently smooth. Reliability of cross-validation for finite samples is discussed.

Key Words: Cross-validation; Smoothing parameters; Kernel density estimation; Histograms

1. Introduction

1.1. Motivation and Scope of Paper

Much theoretical progress has been made recently with the important problem of data-based methods for choosing smoothing parameters in nonparametric curve estimation procedures since the early work of Kronmal and Tarter (1968), Woodroofe (1970), and Stone (1974). In density estimation particular attention has been paid to the least-squares cross-validation algorithm described independently by Rudemo (1982) and Bowman (1984). The sequence of smoothing parameters produced by this procedure not only leads to consistent density estimates but is asymptotically optimal in a certain sense, as shown by Hall (1983) and Stone (1984). Recently Hall and Marron (1985) have characterized the limiting distribution of this sequence. This theory indicates that these cross-validation (CV) sequences converge at perhaps surprisingly slow rates. As was the case with the original kernel theory of Rosenblatt and Parzen, the new theory is asymptotic in nature, so that considerable effort will be required to fully understand the practical aspects of these methods and their performance with real data. For samples of size under 100 with Gaussian kernel estimates, two simulation studies have been completed. First, Scott and Factor (1981) showed that the average behavior of some earlier CV algorithms was good for Gaussian data in the sense that the cross-validation smoothing parameters were centered around the value predicted by minimizing mean integrated squared error ($MISE$). Second, Bowman (1985) presented a study using 6 sampling densities and 8 cross-validation algorithms. We are unaware of studies involving much larger samples.

The goal of cross-validation is to automatically provide nearly optimally calibrated nonparametric estimates, mimicking the choices of experts and perhaps surpassing them. Consistency of cross-validation algorithms is important but we are more concerned with understanding small-sample reliability, which we define as the smallest sample size for there is a 90% chance of being within 10-15% of the optimal smoothing parameter. This is a useful rule of thumb because even for extremely large samples, density estimates with smoothing parameters outside this narrow range are either distorted or visually noisy. Highly reliable cross-validation algorithms would provide scientific reproducibility of density estimates between laboratories, an important but elusive goal. Our objectives are similar to those of researchers trying to retain the reproducibility of multiple linear regression while introducing transformations and robust methods via artificial intelligence (Gale and Pregabon 1983). Carroll and Ruppert (1985) have expressed caution about using robust methods "blindly." We will show that similar caution is appropriate for cross-validation of density estimators.

In this paper we introduce a (biased) cross-validation method closely related to one investigated in Scott and Factor (1981), and compare it with the least-squares unbiased cross-validation algorithm. We remark that our biased cross-validation algorithm is really a generalized rather than ordinary cross-validation method (Wahba 1977). Both theoretical results and simulations indicate that the new algorithm compares favorably to the least-squares cross-validation algorithm in many situations. The theoretical results explain some of the small sample behavior of several cross-validation functions. We also show that cross-validation algorithms can be unreliable for samples sizes which are "too small." Our purpose is to present material that will aid the practitioner in the use of these appealing automatic cross-validation algorithms and to help facilitate evaluation of future algorithms. In order to do this we must address some ofttimes controversial issues in density estimation: squared loss, the integrate squared error and mean integrated squared error criteria, adaptive density estimates, sample size requirements, and assumptions about the underlying density's smoothness. Our simulations include samples of up to 25,600 points.

There is a rich literature on data-based smoothing algorithms for nonparametric methods. A survey of smoothing methods for density estimation may be found in Scott (1986). A more general survey was given by Titterington (1985). In our discussion we shall focus on density estimation, although the situation for nonparametric regression is parallel (Rice 1984; Härdle and Marron 1985). With regression, one must pay attention to the interactions among choices of the regression curve, the signal-to-noise ratio, and the distribution of the noise, whereas we only need consider the density curve here.

At this point it is helpful to give a partial outline of the paper. It would be natural at a first reading to proceed to Sections 5 and 6 after reading Sections 1.2 and 2 and only glancing at theoretical results in Sections 3 and 4. All of the proofs in the paper are presented in Section 10.

A few notations are used extensively throughout the paper. We shall denote the squared L_2 norm of a function ψ by

$$R(\psi) \equiv \| \psi \|_2^2 = \int_{-\infty}^{\infty} \psi(x)^2 dx \ , \tag{1.1}$$

where R reminds us that (1.1) is one possible measure of the roughness of ψ. The square of the p-th derivative of ψ will be denoted by $\psi^{(p)}(x)^2$. Integrals without limits are assumed to be over the entire real line.

1.2. Example

We begin by presenting an example using two cross-validation functions for the histogram. For an equally spaced histogram of a random sample of size n with bin width h, let $\nu_h(k)$ be the bin count in the k-th bin $[kh, (k+1)h)$, where without loss of generality we may assume the mesh includes zero. In Section 3.1, we show that the least-squares cross-validation function is

$$e_0(h) = \frac{2}{nh} - \frac{1}{n^2h} \sum_{k=-\infty}^{\infty} \nu_h(k)^2 \tag{1.2}$$

and, in Section 3.2, propose a biased cross-validation function

$$e_1(h) = \frac{5}{6nh} + \frac{1}{12n^2h} \sum_{k=-\infty}^{\infty} [\nu_h(k+1) - \nu_h(k)]^2. \tag{1.3}$$

The (automatic) cross-validation smoothing parameter minimizes the sample cross-validation function. In Figure 1, we plot e_0 and e_1 for a relatively large sample of 10,000 standard normal points (actually $N(5,1)$), for which the asymptotic L_2 theory (Scott 1979) predicts $h=.162$ minimizes the mean integrated squared error. The difference between these plots is striking. To be sure, most of the "vertical" noise in these plots is due to a bin edge effect. This phenomena was observed even with much smaller samples by Rudemo (1982). But the difference in noise levels has deeper implications. We claim these pictures reveal a great deal about theoretical and practical behavior of these cross-validation techniques for reasonable sample sizes and suggest differences in the "horizontal" noise of smoothing parameters obtained by the two cross-validation methods. Roughly speaking, in Figure 1b we are seeing the between-sample "vertical" variation because of the relatively small correlation between heights of adjacent or partially overlapping bins. An effort to understand these plots was the motivation for this paper.

2. Asymptotic Mean Integrated Squared Error Theory

Consider a kernel density estimate of an unknown univariate density f based on a random sample $x_1, \cdots, x_n$ with corresponding empirical cdf F_n:

$$\hat{f}(y) = \frac{1}{n}\int K_h(x,y)dF_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x_i,y),$$

indexed by a smoothing parameter h. Our goodness-of-fit criterion between f and $\hat{f}$ will be the usual integrated squared error (ISE):

$$ISE = \int_{-\infty}^{\infty} [\hat{f}(y) - f(y)]^2 dy. \tag{2.1}$$

Let $MISE \equiv E(ISE)$, the mean integrated squared error.

We shall focus our attention on the fixed bandwidth (symmetric) kernel estimator

$$\hat{f}\,(y) = \frac{1}{nh}\sum_{i=1}^{n} K\,(\frac{y-x_i}{h}). \tag{2.2}$$

Denote by μ_k the "moments" of the kernel K: $\mu_k = \int t^k K(t)dt$. If we suppose $\mu_0=1$, $\mu_1= \cdots =\mu_{p-1}=0$, and $0< \mid \mu_p \mid <\infty$ for some even p, then the mean integrated squared error may be written as

$$MISE\,(h) = AMISE\,(h) + O\,(n^{-1}, h^{2p+1}), \tag{2.3}$$

where the dominant term is the asymptotic *MISE* (*AMISE*) given by

$$AMISE\,(h) = \frac{R\,(K)}{nh} + (p\,!)^{-2} h^{2p} \mu_p^2 R\,(f^{(p)}), \tag{2.4}$$

where $R\,(K)$ is defined in (1.1). Expression (2.3) holds if we assume that $R\,(K)<\infty$, $f^{(p)}$ is absolutely continuous and $R\,(f^{(p+1)})<\infty$ (Scott 1985). Slightly weaker assumptions on f would give error $o\,(h^{2p})$, but this would render the $AMISE\,(h)$ expression less useful in practice for small samples because the error terms may vanish quite slowly. The *AMISE* is minimized when $h^* = O\,(n^{-1/(2p+1)})$, for example with $p=2$, by

$$h^* = \{R\,(K)/[n\,\mu_2^2 R\,(f'')]\}^{1/5}. \tag{2.5}$$

We will be comparing several smoothing parameters and we adopt the following easily recalled notation: for a particular sample, h_{MISE} minimizes *MISE*, h^* minimizes *AMISE*, $\tilde{h}_{ISE}$ minimizes *ISE*, and $\tilde{h}_{UCV}$ and $\tilde{h}_{BCV}$ minimize the unbiased and biased cross-validation functions. Notice that the last three smoothing parameters depend upon the data.

In this paper we focus on nonnegative kernel estimators, which is the case $p=2$. For our cross-validation results, conditions on the kernel and density will be slightly stronger than those given above. Here we list several sets of conditions, first for the density f and then for the kernel:

($C\,1$): f''' absolutely continuous and f^{iv} integrable; $R\,(f^{iv}\sqrt{f}\,)$ and $R\,(f\,\sqrt{f^{iv}}\,)$ finite

($C\,2a$): $K \geq 0$ symmetric on [–1,1]; K' Holder continuous; $\mu_2>0$

($C\,2b$): K'' absolutely continuous on $(-\infty,\infty)$; K''' continuous on $(-1,1)$; $R\,(K''')<\infty$.

Throughout this paper we use the Gaussian kernel and the triweight kernel, which is defined by

$$K(t) = \frac{35}{32}(1-t^2)^3 \, I_{[-1,1]}(t) \, . \tag{2.6}$$

The triweight kernel is the simplest kernel satisfying conditions ($C\,2a\,,b$).

3. Cross-Validation Algorithms and Theory

3.1. Least-Squares (Unbiased) Cross-Validation

Ideally, for each sample, we would like to construct a density estimate to minimize the integrated squared error (2.1). Such a strategy would seem an improvement compared to a strategy that minimizes mean integrated squared error. We comment further on this point in Section 7. The least-squares cross-validation criterion attempts to address integrated squared error rather than mean integrated squared error. We shall introduce the least-squares cross-validation criterion for the generalized adaptive kernel estimator, which includes most commonly used estimators such as the histogram. Replacing $\hat{f}$ by the generalized estimator $\hat{g}$ in (2.1) and expanding yields

$$ISE = \int \hat{g}(y)^2 dy - 2\int \hat{g}(y) f(y) dy + \int f(y)^2 dy \, . \tag{3.1}$$

Here,

$$\hat{g}(y) = \frac{1}{n} \sum_{k=1}^{n} K_{nk}(y\,,x_k), \tag{3.2}$$

where the kernel depends on the sample size n and may also depend on either y or x_k. The idea of Rudemo (1982) and Bowman (1984) is to find data-based expressions that, on average, agree with the first two terms in (3.1), and to omit the third term $R(f)$, which amounts to a simple fixed shift of the entire function for each sample. Consider the cross-validation estimator

$$UCV \equiv \int \hat{g}(y)^2 dy - \frac{2}{n} \sum_{i=1}^{n} \hat{g}_i(x_i) \tag{3.3}$$

where

$$\hat{g}_i(x_i) = \frac{1}{n-1} \sum_{k \neq i} K_{nk}(x_i\,,x_k) \, . \tag{3.4}$$

Notice that the divisor has been changed from n to $n-1$, but this change is not incorporated into the kernel. Now the expectation of UCV exactly matches the $MISE$, which is the expectation of

(3.1), term by term since

$$E\hat{g}_i(x_i) = EK_{nk}(x_i, x_k) = E\int K_{nk}(y, x_k)f(y)dy = E\int \hat{g}(y)f(y)dy \ . \qquad (3.5)$$

Expression (3.5) requires the use of Fubini's theorem, and noting that the expectations are of different dimension.

Hence, in the fixed bandwidth case (2.2), ***exactly*** unbiased estimates of the shifted *MISE* for nonrandom h are provided by

$$UCV(h) = \int \hat{f}(y)^2 - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_i(x_i). \qquad (3.6)$$

We refer to (3.6) as an ***unbiased cross-validation*** (UCV) criterion because its expectation is

$$E\Big[UCV(h)\Big] = MISE(h) - R(f) \ . \qquad (3.7)$$

Other theoretical expressions such as (2.4) are only asymptotically correct; that is, they are biased for finite samples.

It is straightforward to see that (1.2) follows from (3.3), except for replacing $n \pm 1$ by n. Hall (1983) and Stone (1984) have shown that the procedure not only provides a consistent sequence of smoothing parameters but is asymptotically optimal in a certain sense. There are (at least) two other remarkable features of this procedure, namely, its unbiasedness property, and its self-adapting property. To illustrate the second property, consider the fixed kernel estimator (2.2). Proper analysis of its *MISE* in (2.3) required knowledge of the "moments" of the kernel, defined below expression (2.2). Such specification is not apparent in (3.6), which in this case becomes,

$$UCV(h) = \frac{R(K)}{nh} + \frac{1}{n(n-1)}\sum\sum_{i \neq j}\left[\int \frac{1}{h^2}K(\frac{x-x_i}{h})K(\frac{x-x_j}{h})dx - \frac{2}{h}K(\frac{x_i-x_j}{h})\right]. \qquad (3.8)$$

An instructive exercise is to show that for p even:

$$E\left[\int \frac{1}{h^2}K(\frac{x-x_i}{h})K(\frac{x-x_j}{h})dx\right] = \int[\int \frac{1}{h}K(\frac{x-s}{h})f(s)ds]^2 dx$$

$$= R(f) + (p!)^{-2}h^{2p}\mu_p^2 R(f^{(p)}) + 2\sum_{k=p/2}^{p}(-1)^k\frac{1}{(2k)!}h^{2k}\mu_{2k}R(f^{(k)}) + O(h^{2p+2}).$$

and

$$E\left[\frac{1}{h}K(\frac{x_i-x_j}{h})\right] = \int\left[\frac{1}{h}\int K(\frac{s-t}{h})f(s)ds\right]f(t)dt$$

$$= R(f) + \sum_{k=p/2}^{p}(-1)^k \frac{1}{(2k)!}h^{2k}\mu_{2k}R(f^{(k))}) + O(h^{2p+2}),$$

Hence,

$$E[UCV(h)] = \frac{R(K)}{nh} + (p!)^{-2}h^{2p}\mu_p^2 R(f^{(p)}) - R(f) + O(n^{-1}), \qquad (3.9)$$

which is to be compared to (2.3), (2.4), and (3.7). Higher order terms in (3.9) match as well in the *MISE* expansion if the necessary derivatives are assumed to exist. Thus the *UCV* criterion automatically "knows" the correct order of the kernel. (In fact, Stone (1984) has shown that the method even knows how many derivatives f has for the histogram and nonnegative kernel estimators.) Notice that for large sample sizes, the *UCV* essentially provides estimates of $R(f^{(p)})$.

In view of Figure 1, the "vertical" variability of $UCV(h)$ is of interest. Assume now that the kernels are symmetric and have finite support on [–1,1]. If we define

$$\gamma(c) = \int K(w)K(w+c)dw - 2K(c) \qquad (3.10)$$

and let $c_{ij} = (x_i - x_j)/h$, then (3.8) becomes (replacing $n-1$ by n)

$$UCV(h) = \frac{R(K)}{nh} + \frac{2}{n^2h}\sum_{i<j}\sum\gamma(c_{ij}). \qquad (3.11)$$

The following theorem provides the mean and variance of the unbiased cross-validation function (3.11) for fixed h.

Theorem 3.1. *For the unbiased cross-validation kernel criterion* (3.11)

$$E[UCV(h)] = AMISE(h) - R(f) + O(n^{-1}) \qquad (3.12)$$

$$Var[UCV(h)] = \frac{4}{n}[R(f^{3/2}) - R(f)^2] + O(1/n^2h + h^4/n). \qquad (3.13)$$

The variance of the histogram criterion (1.2) *is also given by* (3.13).

We prove this in Section 10.1. The rate $O(n^{-1})$ of the leading term in (3.13) was noted by Rudemo (1982). Observe that as a consequence of Jensen's inequality the first term in (3.13) is nonnegative.

Remark: For the example in Figure 1, $\sqrt{Var}$ =.00222 for $h=h^*$=.162. This noise, indicated by the longer vertical line in the bottom right corner of Figure 1b, is much greater than the observed noise. We shall return to this point in Section 3.3.

3.2. Biased Cross-Validation

The asymptotic expansion for the mean integrated squared error as given in (2.4) contains only one unknown quantity, $R(f^{(p)})$. One natural estimator is $R(\hat{f}^{(p)})$, where $\hat{f}$ is the fixed bandwidth kernel estimator. Scott, Tapia, and Thompson (1977) used this estimator in a fixed point algorithm for choosing h in the case $p=2$. However, the following lemma (proved in Section 10.2) shows that this estimator is deficient asymptotically and indicates how an improved estimator can be constructed.

Lemma 3.2: *Suppose that derivatives of order $p+2$ of the density f and kernel K exist and are continuous. Then*

$$E[R(\hat{f}^{(p)})] = R(f^{(p)}) + \frac{R(K^{(p)})}{nh^{2p+1}} + O(h^2)\,. \tag{3.14}$$

Notice that for smoothing parameters of the optimal order $h_p^* = c_p\, n^{-1/(2p+1)}$, the kernel estimate provides a positively biased estimate of $R(f^{(p)})$, but by an asymptotically *constant* amount. Silverman (1978) based his visual "test graph" method for choosing h on another characterization of this asymptotic bias in the L_∞ norm. An improved estimate of $R(f^{(p)})$ is

$$\hat{R}(f^{(p)}) \equiv R(\hat{f}^{(p)}) - \frac{R(K^{(p)})}{nh^{2p+1}}. \tag{3.15}$$

Special Case $p=2$: For the important case of the nonnegative kernel method when $p=2$, let

$$\phi(c) \equiv \int K''(w)K''(w+c)\,dw\ . \tag{3.16}$$

Then

$$R(\hat{f}'') = \frac{R(K'')}{nh^5} + \frac{2}{n^2h^5}\sum_{i<j}\sum\phi(c_{ij})\,,$$

where (again) $c_{ij}=(x_i-x_j)/h$. Using this together with the correction (3.15) in the *AMISE* expression (2.4) defines a biased cross-validation function:

$$BCV(h) \equiv \frac{R(K)}{nh} + \frac{\mu_2^2}{2n^2h}\sum_{i<j}\sum \phi(c_{ij}) \qquad (3.17)$$

Notice that the two $R(K'')/nh^5$ terms cancel. Observe the similarities between (3.11) and (3.17); both are U-statistics but with different kernels. Expression (3.17) will provide useful estimates of the *MISE*.

Then we have the following interesting results for the biased cross-validation estimator.

Theorem 3.2: *For a nonnegative kernel estimator satisfying conditions* $(C\,1)$ *and* $(C\,2b)$, *the estimator* $BCV(h)$ *is asymptotically normal with mean and variance*

$$E[BCV(h)] = AMISE(h) + O(n^{-1}) \qquad (3.18)$$

$$Var[BCV(h)] = \mu_2^4\, R(\phi) R(f) \,/\, (8n^2h) + O(h/n^2). \qquad (3.19)$$

For the histogram cross-validation estimator given by (1.3),

$$Var[e_1(h)] = R(f)/(12n^2h) + O(n^{-2}). \qquad (3.20)$$

For $h = O(n^{-1/5})$, $Var = O(n^{-9/5})$ in (3.19). It follows from (3.18) and (2.3) that the bias in $BCV(h)$ is $O(n^{-1})$. Thus the squared bias is $O(n^{-2})$, which is of lower order than the variance by the factor h. Hence variance dominates "vertical" mean squared error. Note that the results of Theorem 3.1 and Theorem 3.2 are not to comparable orders, since $Var = O(n^{-1})$ in (3.13). This discrepancy is resolved in the next section. From (3.20) we may compute $\sqrt{Var}$ =.0000381 at h^*=.162, which closely approximates the observed variation in Figure 1a, as indicated by the small vertical line in the bottom right corner.

It follows from this theorem that a consistent sequence of smoothing parameters can be found.

Corollary 3.2: *Let* $\tilde{h}_{BCV}$ *minimize* (3.17) *over* $(0,bh^*)$ *for any* $b>1$. *Then*

$$\underset{n\to\infty}{plim}\,(\tilde{h}_{BCV}/h^*) = 1 \qquad (3.21)$$

3.3. Unbiased Cross-Validation Revisited

3.3.1. Augmented Unbiased Cross-Validation Criterion

The reason that the variation computed in Theorem 3.1 is not comparable to that of Theorem 3.2 is that Theorem 3.1 measures the vertical variation of the *UCV* curve about the level *MISE* $-R(f)$ rather than the *MISE* level, which is converging to 0. The vertical variation of the entire curve has no effect on the location of the minimum we are interested in. Bowman's (1984) method of derivation gave the following augmented $UCV(h)$ formula, which Hall (1983) argued is the correct form for theoretical analysis:

$$AUCV(h) = R(\hat{f}) - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_i(x_i) + \frac{2}{n}\sum_{i=1}^{n} f(x_i) - R(f) \qquad (3.22)$$

With this change, Theorem 3.1 is revised to give variance results of the same order as that in Theorem 3.2.

Theorem 3.3: *For nonnegative kernel estimators satisfying conditions* $(C1)$ *and* $(C2a)$,

$$Var[AUCV(h)] = 2R(\gamma)R(f)/(n^2h) + O(h/n^2). \qquad (3.23)$$

For the histogram, we augment (1.2) *and find*

$$Var\left[e_0(h) + \frac{2}{n}\sum_{i=1}^{n} f(x_i) - R(f)\right] = h^2R(f'\sqrt{f})/n + 2R(f)/(n^2h) + o(n^{-5/3}). \qquad (3.24)$$

Corollary 3.3: *Let* $\tilde{h}_{BCV}$ *minimize (3.22) or equivalently (3.11) over* (ah^*, bh^*) *for arbitrarily small* a *and large* b. *Then*

$$\underset{n\to\infty}{plim}\,(\tilde{h}_{UCV}/h^*) = 1 .$$

From (3.24) we compute $\sqrt{Var}$ =.0003394 at h^*=.162, indicated by the smaller vertical line in the bottom right corner of Figure 1b. This closely approximates the observed variation. However, this standard deviation is 8.9 times larger than for the biased criterion in Figure 1a. Conditions $(C1)$ are much stronger than those required by Hall and Marron (1985) due to our different approach.

3.3.2. Asymptotic Relative ``Vertical'' Variability

It is now a simple matter to compare Theorems 3.2 and 3.3 for kernel estimates. The relative variability may be defined by the ***square root*** of the ratio of the variances (3.23) and (3.19):

$$ratio = \frac{4}{\mu_2^2}\left[\frac{R(\gamma)}{R(\phi)}\right]^{1/2}, \qquad (3.25)$$

which only depends on the kernel. This ratio exceeds 10 for most kernels. In the first part of Table I, we have computed this ratio for several practical kernels.

The extent to which the "vertical" noise is converted to "horizontal" noise is examined empirically in Section 6 and theoretically in Section 4.

4. Variability and Asymptotic Normality of CV Smoothing Parameters

In Section 3 we characterized the "vertical" variability of the cross-validation functions. Of more practical interest is the "horizontal" variability of the actual cross-validation estimates $\tilde{h}_{UCV}$ and $\tilde{h}_{BCV}$. In terms of ratios, we show that about half of the "vertical" error is translated into "horizontal" error.

4.1. Unbiased Cross-Validation

Hall and Marron (1985) investigate the variability of $\tilde{h}_{UCV}$ about the idealized target $\tilde{h}_{ISE}$ and show that $\tilde{h}_{UCV}-\tilde{h}_{ISE}$ is asymptotically normal (AN). Because (as we will see in Sections 6 and 7) $\tilde{h}_{UCV}$ and $\tilde{h}_{ISE}$ are often negatively correlated, we now compute the variation of $\tilde{h}_{UCV}$, or equivalently, of $\tilde{h}_{UCV}-h^*$. We examine the first derivative of $UCV(h)$ given in (3.11), since the extra terms in the augmented criterion (3.22) do not involve h. Let $\gamma_+(c)$ and $\gamma_-(c)$ define $\gamma(c)$ given in (3.10) on the intervals [0,2] and [–2,0], respectively:

$$\gamma_+(c) \equiv \int_{-1}^{1-c} K(w)K(w+c)dw - 2K(c) \qquad 0 \le c \le 2 \qquad (4.1)$$

and $\gamma_-(c)$ as in (4.1) with limits $-1-c$ and 1. Consider the derivative of $\gamma_+(c_{ij})$:

$$\frac{d}{dh}\gamma_+(c_{ij}) = \frac{-c_{ij}}{h} \int_{-1}^{1-c_{ij}} K(w)K'(w+c_{ij})dw + \frac{2c_{ij}}{h} K'(c_{ij}) \quad 0 \le c_{ij} \le 2 ,$$

where the other term involving the derivative of the upper endpoint in the integral vanishes since $K'(1)=0$. If we define

$$\rho(c) = c \int K(w)K'(w+c)dw - 2cK'(c) \quad -2 \le c \le 2 \tag{4.2}$$

and zero elsewhere, then $\tilde{h}_{UCV}$ satisfies

$$\frac{d}{dh} UCV(h) \Big|_{h=\tilde{h}_{UCV}} = 0$$

or, equivalently,

$$\sum_{i<j}\sum [\gamma(c_{ij}) + \rho(c_{ij})] \Big|_{h=\tilde{h}_{UCV}} = -nR(K)/2 . \tag{4.3}$$

Hall and Marron have shown that the left-hand side (which is a degenerate martingale) is AN. In Section 10.5 we compute the moments and find

Lemma 4.1: *Under conditions (C1) and (C2a),*

$$\sum_{i<j}\sum [\gamma(c_{ij}) + \rho(c_{ij})] = AN\{-n^2h^5\mu_2^2R(f'')/2 , n^2hR(\rho)R(f)/2\} \tag{4.4}$$

Now $\underset{n\to\infty}{plim} (\tilde{h}_{UCV}/h^*) = 1$ so that we may replace $\tilde{h}_{UCV}$ by h^* in the variance. Hence (4.3) becomes

$$-n^2\tilde{h}_{UCV}^5\mu_2^2R(f'')/2 = AN\{-nR(K)/2 , n^2h^*R(\rho)R(f)/2\} . \tag{4.5}$$

Dividing we have

$$\tilde{h}_{UCV}^5 = AN\{R(K) / [n\mu_2^2R(f'')] , 2h^*R(\rho)R(f) / [n^2\mu_2^4R(f'')^2]\} . \tag{4.6}$$

But the mean is simply $(h^*)^5$ by (2.5). Hence

$$(\tilde{h}_{UCV}/h^*)^5 = AN\{1 , 2R(\rho)R(f) / [n^2(h^*)^9\mu_2^4R(f'')^2]\} . \tag{4.7}$$

Since the variance $\to 0$ as $n \to \infty$, we may apply the delta method (Serfling 1980) with $g(x)=x^{1/5}$, which reduces the variance by the factor 25. Multiplying through by h^*, we have

Theorem 4.1: *For a nonnegative kernel estimator satisfying conditions (C1) and (C2a),*

$$\tilde{h}_{UCV} = AN\{h^* , 2R(\rho)R(f) / [25n^2h^{*7}\mu_2^4R(f'')^2]\} . \tag{4.8}$$

Now set $h^* = c_2n^{-1/5}$ *; then the standard deviation is given by*

$$\sigma(\tilde{h}_{UCV} - h^*) = \frac{\sqrt{2}\, c_2^{-7/2}}{5\mu_2^2 R(f'')} \sqrt{R(\rho) R(f)}\; n^{-3/10} . \tag{4.9}$$

The relative error of $\tilde{h}_{UCV}$ *is* $O(n^{-1/10})$.

4.2. Biased Cross-Validation

In a similar fashion we may investigate the limiting distribution of $\tilde{h}_{BCV} - h^*$. Define

$$\psi(c) = c \int K''(w) K'''(w+c)\, dw \qquad -2 \leq c \leq 2 \tag{4.10}$$

and zero elsewhere. Then taking the derivative of (3.17), we find

$$\sum_{i<j}\sum [\phi(c_{ij}) + \psi(c_{ij})] \Big|_{h = \tilde{h}_{BCV}} = -2nR(K) / \mu_2^2 . \tag{4.11}$$

In Section 10.6 we compute the first two asymptotic moments (AM) of (4.11) and obtain

Lemma 4.2: *Under conditions (C1) and (C2a,b),*

$$\sum_{i<j}\sum [\phi(c_{ij}) + \psi(c_{ij})] = AM\{-2n^2h^5R(f''),\, n^2hR(\psi)R(f)/2\} \tag{4.12}$$

Again $\underset{n \to \infty}{plim}\, (\tilde{h}_{BCV} / h^*) = 1$, so that we may use h^* in the variance. In a direct fashion we find

$$\tilde{h}_{BCV}^5 = AM\{(h^*)^5,\, h^* R(\psi) R(f) / [8n^2 R(f'')^2]\} . \tag{4.13}$$

Applying the delta method as before,

Theorem 4.2: *Under the conditions of Lemma 4.2*

$$\tilde{h}_{BCV} = AM\{h^*,\, R(\psi)R(f) / [200n^2 (h^*)^7 R(f'')^2]\} \tag{4.14}$$

and

$$\sigma(\tilde{h}_{BCV} - h^*) = \frac{c_2^{-7/2}}{10\sqrt{2} R(f'')} \sqrt{R(\psi) R(f)}\; n^{-3/10} . \tag{4.15}$$

We remark that we believe it can be shown that $\tilde{h}_{BCV}$ is AN in (4.14).

4.3. Asymptotic Relative "Horizontal" Variability

Comparing the standard deviations in (4.9) and (4.15), we see that the asymptotic relative "horizontal" efficiency defined as the ratio of these standard deviations (not variances) is

$$ratio = \frac{4}{\mu_2^2}\sqrt{\frac{R(\rho)}{R(\psi)}} \,. \tag{4.16}$$

Compare this to the ratio for the "vertical" noise given in expression (3.25). For the triweight kernel this ratio is 4.98; see Table I. The usefulness of these results in practice is discussed in the remainder of the paper.

5. Implementation with Gaussian Kernel and Averaged Shifted Histogram Estimators

5.1. Two Introductory Examples

The development thus far requires kernels with finite support. However, it extends to kernels with exponentially decreasing tails as is the case with the Gaussian kernel. For this important case, which was considered by Rudemo (1982) and Bowman (1984), equations (3.11) and (3.17) become

$$UCV(h) = \frac{1}{2\sqrt{\pi}nh} + \frac{1}{\sqrt{\pi}n^2h}\sum_{i<j}\sum \left(e^{-c_{ij}^2/4} - 2\sqrt{2}e^{-c_{ij}^2/2} \right), \tag{5.1}$$

and

$$BCV(h) = \frac{1}{2\sqrt{\pi}nh} + \frac{1}{64\sqrt{\pi}n^2h}\sum_{i<j}\sum (c_{ij}^4 - 12c_{ij}^2 + 12)e^{-c_{ij}^2/4}, \tag{5.2}$$

respectively. (Equation (5.1) actually replaces terms like $n+1$ with n.) We plot (5.1) and (5.2) in Figures 2a and 2b for samples of size 25 and 400 from $N(0,1)$, using data generated by IMSL routine GGNPM with seeds 1821291829 and 1943248741, respectively. For plotting purposes, we have augmented $UCV(h)$ as in equation (3.22). The dotted line represents the (exact) $MISE$ as a function of h (Fryer 1976), which is given by

$$MISE(h,n) = \frac{1}{2\sqrt{\pi}nh} + \frac{n-1}{2n\sqrt{\pi(1+h^2)}} - \frac{2}{\sqrt{2\pi(2+h^2)}} + \frac{1}{2\sqrt{\pi}} \,.$$

For fixed n the BCV function converges to 0 as $h \to \infty$. The BCV function barely exhibits a local minimum with $n = 25$ (sometimes it has none; see Section 6), but exhibits a clear local minimum when $n = 400$. Heuristically, the BCV indicates the quality of $\tilde{h}_{BCV}$ by the amount of rise to the right of the minimum. As n increases, BCV provides reasonable estimates of $MISE$

for relatively larger values of h beyond h_{MISE}. Recall that *MISE* is increasingly dominated by bias in the region to the right of h_{MISE}.

The *UCV* function does relatively well in the high bias region and less well in the high variance region, which is to the left of h_{MISE}, as predicted by Theorem 3.3. There is no high frequency component evident in individual plots as was the case for the histogram in Figure 1a since we are using a continuous kernel. With $n = 400$ we have selected a case where the *UCV* function has a minimum well to the left of h_{MISE}; see Section 6.1. (The minima in Figure 2b are 0.142, 0.330, and 0.389). Rudemo in a draft of his 1982 paper observed this (occasional) behavior for smaller samples and speculated it was consistent with features in the data. In Figure 2c we plot the two *CV* estimates along with the true density (the h_{MISE} estimate is quite similar to the *BCV* estimate). The density estimate reveals the illusory multimodal feature that attracted the *UCV* function. In Figure 2b, observe the inflection points in both *CV* curves in the neighborhood of the other's minimum. The *UCV* (5.1) also eventually converges to 0 (the augmented version to approximately $R(f)$); the curve in Figure 2b increases monotonically to $-.579$ on the log scale.

For $n = 400$, evaluating (5.1) and (5.2) for each h took more than 1.1 CPU minutes on a VAX 11/750. Figure 2b required several hours of CPU. Clearly an alternative implementation is required for even moderate sample sizes.

5.2. Averaged Shifted Histogram Implementation

In order to carry out an extensive Monte Carlo study, it is necessary to find a more computationally feasible method than the very slow Gaussian kernel implementation given above. Much faster evaluations of (3.11) and (3.17) are possible with finite support kernels. Furthermore, a kernel procedure using binned data accelerates *CV* algorithms even more, for example, Silverman's (1982) Fast Fourier Transform algorithm. Another procedure that takes advantage of binned data is the averaged shifted histogram (ASH) (Scott 1985). An ASH is the (weighted) average of m histograms, each with bin width h but with bin mesh origins at integer multiples of

$\delta \equiv h/m$, and is given by:

$$\hat{f}_m(y) = \frac{1}{nh} \sum_{i=1-m}^{m-1} w_m(i)\nu_\delta(k+i) \quad \mathit{for}\ y\ \mathit{in}\ I_k\ , \tag{5.3}$$

where $w_m(i)$ are the weights and $\nu_\delta(k)$ is the bin count for the k-th bin $I_k \equiv [k\delta,(k+1)\delta)$. The weights corresponding to the triweight kernel (2.6) are

$$w_m(i) = c_m\,[1-(i/m)^2]^3 \quad \text{for } |i| < m\ , \tag{5.4}$$

where c_m is a normalizing constant so that $\sum w_m(i) = m$ given by

$$c_m = 35\ /\ [32(1-1/4m^2)(1+1/4m^2+5/24m^4)]\ .$$

The UCV formula (3.3) for $\hat{f}_m$ is easily evaluated. The term $\int \hat{f}_m(y)^2 dy$ is computed directly. The term $\sum_i \left(\hat{f}_m\right)_i(x_i)$ in (3.3) and (3.4) is simply equal to

$$\sum_{k=-\infty}^{\infty} \nu_\delta(k)\ \hat{s}_k\ -\ \frac{w_m(0)}{h}\ , \tag{5.5}$$

where $\hat{s}_k \equiv \hat{f}_m(k\delta)$ is the value of $\hat{f}_m$ in I_k. In practice the sum in (5.5) involves perhaps a few hundred terms. For $m>10$ (i.e. δ sufficiently small) the behavior of the kernel and ASH estimators is virtually identical; in particular, similar values of the smoothing parameter h give nearly identical results.

For BCV, the asymptotic theory for the ASH involves both $R(f')$ and $R(f'')$, which is unfortunate. However, the frequency polygon (linear interpolator) of the ASH (FP-ASH) requires only $R(f'')$. We cannot use binned data with UCV on FP-ASH since we would need to know $\hat{g}_i(x_i)$ - i.e. need to know all the x_i exactly and not just $\hat{s}_k$ (or equivalently, the bin counts) as in ASH case. Again we emphasize that for $m>10$, the ordinary kernel, ASH, and FP-ASH are essentially the same for the same h.

The asymptotic $MISE$ expression for the FP-ASH is (Scott 1985)

$$\mathit{AMISE} = \frac{2R_w + \gamma_w}{3nh} + \frac{1}{4}h^4\left[\sigma_w^4 + \frac{\sigma_w^2}{2m^2} + \frac{49}{720m^4}\right]R(f'')$$

where $mR_w = \sum w_m(i)^2$, $m\gamma_w = \sum w_m(i)w_m(i-1)$, and $m^3\sigma_w^2 = \sum i^2 w_m(i)$. Our estimate of $R(f'')$ turns out to be:

$$\hat{R}(f'')=\frac{1}{\delta^3}\sum_k(\hat{s}_{k+1}-2\hat{s}_k+\hat{s}_{k-1})^2-\frac{m^3}{nh^5}\left[2w_{m-1}^2+4(w_0-w_1)^2+2\sum_{i=1}^{m-1}(w_{i+1}-2w_i+w_{i-1})^2\right] \quad (5.6)$$

where we denote $w_m(i)$ by w_i and $\hat{s}_k$ is defined below equation (5.5). These may be computed in closed form using MACSYMA (the triweight kernel formulae are available from us).

6. Monte Carlo Study

6.1. Small-to-Large Sample Behavior with Gaussian Data

In this section we study the results of simulations based on samples from a standard normal distribution for sample sizes $n=$ 25, 100, 400, 1600, 6400, and 25600 with repetitions of 250, 200, 150, 100, 100, and 100, respectively. ASH and FP-ASH estimators with a triweight kernel were used as described above. The δ's chosen were .15, .10, .05, .025, .02, and .01, respectively. For each sample, *ISE* 's corresponding to 4 different bandwidths h (or equivalently m since $h=m\,\delta$) were computed numerically: h_{MISE}, $\tilde{h}_{ISE}$, $\tilde{h}_{BCV}$, and $\tilde{h}_{UCV}$. The value $\tilde{h}_{ISE}$, which minimizes the *ISE* for a particular sample, was found by searching over integer values m .

In Figure 3 we plot frequency polygons of the cross-validated smoothing parameters. The vertical lines indicate the set of h 's examined (multiples of δ). In Table II we present some summary statistics. We note immediately that in 103/250 samples with $n=25$, the *BCV* function had no local minima (compare Figure 2a). The average of $\tilde{h}_{UCV}$ when $n=25$ is reasonable, but only a relatively few individual samples are close to h_{MISE}. (Of course, perhaps $\tilde{h}_{ISE}$ isn't close. We check this below.) We have not found any samples where the *BCV* failed to have a local minimum for $n>40$. (For other densities this threshold is higher.) On the other hand, the biased *CV* estimates tighten up quickly beyond this threshold so that the "worst" case for $n\geq1600$ is quite close to h_{MISE} (a reasonable target as we discuss in Section 7). The unbiased *CV* procedure continues to be attracted to spurious (rough) estimates even with $n=25600$. Its convergence to normality is also apparently slower. The asymptotic theory predicts a ratio of "vertical" standard errors of the *CV* curves of 11.65 (which was observed in the simulations) and a ratio of "horizontal" standard errors of *CV* smoothing parameters of 4.98; see Table I. In Table II we see that the finite sample ratio is reasonably close to 4.98 for moderate sample sizes and that expressions (4.15) and (4.9), which yield $\sigma(\tilde{h}_{BCV}) = .250n^{-3/10}$ and $\sigma(\tilde{h}_{UCV}) = 1.243n^{-3/10}$, are remarkably accurate.

A more detailed study of the individual results for n =400 and n =25600 is worthwhile. In Figure 4 we compare the various smoothing parameters. Surprisingly, there is a negative correlation between $\tilde{h}_{ISE}$ and $\tilde{h}_{UCV}$ (−.41 and −.38, respectively); see Section 7. The $\tilde{h}_{BCV}$ clusters more tightly around h_{MISE} (the correlations with $\tilde{h}_{ISE}$ are −.44 and −.16, respectively). For the 150 repetitions with n=400, 41 had $\tilde{h}_{UCV} \leq .85$ (.85 was the smallest observed $\tilde{h}_{BCV}$ value). In 23 of 150 samples the *UCV* curve had 2 minima, always one less and one greater than .85. Seven of these had a more reasonable local minimum near h^*. Sixteen (all $\leq .85$) were local minima compared to a reasonable $\tilde{h}_{UCV}$ near h^*. When n=25600, only 2 of 100 *UCV* curves had a second (local) minimum, but in both cases the global minimizer was more reasonable. None of the *BCV* curves had any other local minima over the range searched.

In Figure 5 the numerically computed *ISE* 's of these samples are displayed. From Figures 5a and 5d, h_{MISE} is only occasionally grossly inefficient relative to $\tilde{h}_{ISE}$. In Figures 5b and 5e we see that the *BCV* almost dominates the *UCV* estimates with respect to *ISE* ! We try to understand the *BCV* estimates in Section 7. Figures 5c and 5f are presented for completeness.

Using the Hall and Marron (1985) formulae for the triweight kernel and Gaussian data, we obtain $\sigma(\tilde{h}_{ISE}) = 1.304n^{-3/10}$ and $\sigma(\tilde{h}_{UCV} - \tilde{h}_{ISE}) = 2.081n^{-3/10}$. Since $\sigma(\tilde{h}_{UCV}) = 1.243n^{-3/10}$, it follows that there is indeed a negative correlation between $\tilde{h}_{UCV}$ and $\tilde{h}_{ISE}$. With the data above, we computed the sample version of $\sigma(\tilde{h}_{UCV} - \tilde{h}_{ISE})$ as .3464 and .0918 for n =400 and n =25600, respectively, which agree closely with the theoretical predictions of .3448 and .0990. Thus while the variability of $\tilde{h}_{UCV}$ and $\tilde{h}_{ISE}$ is similar, the negative correlation suggests they are often on opposite sides of h_{MISE}. We have seen how $\tilde{h}_{BCV}$, which is very close to h_{MISE} for large samples, generally corresponds to estimates with integrated squared errors smaller than using $\tilde{h}_{UCV}$.

6.2. Other Densities

Similar simulations were performed for three other densities: Cauchy, Lognormal (exponential of standard Gaussian random variable), and a mixture given by

$$f\ (x\ ) = .75\ \phi(x\ ;0,1) + .25\ \phi(x\ ;2,\frac{1}{9})$$

where $\phi(x\ ;\mu,\sigma^2)$ is the normal density with mean μ and variance σ^2. Each of these densities can be adequately estimated by a fixed bandwidth estimator, but an appropriate variable bandwidth estimator would be useful for small samples. The efficiency of a fixed bandwidth nonnegative kernel estimator relative to a variable bandwidth nonnegative kernel estimator with $h\ = h_x$ (that is, a different h for every point estimate $f\ (x\ )$) may be seen to be

$$\frac{AMISE\,(adaptive\,)}{AMISE\,(f\,ixed\,)} = \frac{\int \left[f''(x\,)^2 f\ (x\,)^4 \right]^{1/5} dx}{\left[\int f''(x\,)^2 dx \right]^{1/5}} . \tag{6.1}$$

This ratio is equal to .915, .767, .640, and .728, for the Gaussian, Cauchy, Lognormal, and Mixture densities, respectively.

In Figure 6 we plot histograms of the CV estimates for 100 repetitions with n =1600. The Cauchy simulations ($25 \leq n \leq 25600$) were similar to the Gaussian results in Section 6.1 except that 17% of the BCV estimates failed to exist for n =100 and the ratios of "horizontal" standard errors increased to only 3.0; see Table III.

The lognormal and mixture results have an interesting twist. Notice the BCV estimates are shifted to the right from the UCV estimates. BCV failures were observed at n =400. The UCV continued to perform as usual: average behavior close to h_{MISE} with high variability. The BCV was definitely biased upward for moderate sample sizes (not including samples where BCV did not exist), although the bias vanishes by n =25600. We understand this phenomenon as follows: the estimates which are optimal with respect to ISE are relatively rough or noisy. This is not a defect of L_2 error but of the fixed bandwidth estimator; see Section 7.

We examine two particular examples. In Figure 7, density estimates of a lognormal sample with n=1600 are shown. We used δ=.015 and cross-validated over the interval (-1,14). With this sample $h=.21$ $(m=14)$ for the *UCV*, *ISE*, and *MISE* criteria while $\tilde{h}_{BCV}=.33$ $(m=22)$. The minimum *ISE* estimate in Figure 7b is quite rough, even near the mode. The *BCV* estimate in Figure 7a smoothes more appropriately near the mode and has somewhat reduced noise in the tail (its *ISE* is 34% larger). It is interesting to note that 90% of $R(f'')$ comes from the interval (0,.0257)! (We remark that an adaptive estimator would not be much improved near the mode but rather in the tail.) L_1 and L_∞ errors are minimized for $h=.225$ and .165, respectively.

Figure 8 is a plot of kernel estimates of a mixture sample with $n=400$. Again $\delta=.015$. For this sample, $h_{MISE}=.615$, $\tilde{h}_{ISE}=.510$, $\tilde{h}_{UCV}=.480$, and $\tilde{h}_{BCV}=.870$ (with *ISE* greater by 55%). L_1 and L_∞ errors are minimized for $h=.525$ and .540, $(m=35$ and $36)$, respectively. The *UCV* estimator is best in the narrow peak while the *BCV* is better in the larger peak. For a fixed bandwidth estimator, we prefer the estimate in Figure 8a (consistent with earlier recommendations of Fryer (1976) to slightly exceed h^*). On the other hand for small samples, occasionally large $\tilde{h}_{BCV}$'s are produced that obscure the bimodal feature.

7. Some Issues in Cross-Validation

One issue that reoccurs is whether we should use with a smoothing parameter that minimizes *MISE* (or *AMISE*) or whether we should minimize the *ISE* for the data at hand. In theory, we should address *ISE*. Two factors primarily affect the (optimal) *ISE* of a density estimate for individual samples. The first is our use of a fixed bandwidth estimator when a variable bandwidth estimator is more appropriate; see equation (6.1). We do not address this subject in any more detail, except to note that its cross-validation is more delicate due to an increase in number of smoothing parameters. The second factor is variation in the lower order sample moments. While no choice of smoothing parameter can compensate for a shift in mean, it is possible to reduce *ISE* due to variation in the sample variance. If $\hat{\sigma}<\sigma$, then choose $h>h^*$ and vice versa. In practice we cannot expect a cross-validation method to successfully mimic the behavior of $\tilde{h}_{ISE}$. To do so would require guessing whether $\hat{\sigma}>\sigma$ or vice versa. Clearly this requires knowledge about the unknown density. Thus we do not expect much, if any, improvement for *CV* methods attempting to minimize *ISE* compared to those seeking the single fixed bandwidth minimizing *MISE* as we saw in Section 6. Bowman (1985) found that using a simple rule such as $h = \hat{\sigma} n^{-1/5}$ worked (embarrassingly) well with respect to *ISE* for a Gaussian kernel estimator. Such a rule is close to a *MISE* rule for many unimodal densities.

It is easy to demonstrate these observations by simulation. We chose standard Gaussian data since fixed bandwidth estimates are 91.5% efficient. We used the moderate and large sample sizes of 400 and 25,600 with 150 and 100 repetitions, respectively, of Section 6. In Figure 9 we plot $\tilde{h}_{ISE}$ vs $\hat{\sigma}$ for sample sizes 400 and 25,600. The correlations between $\tilde{h}_{ISE}$ and $\hat{\sigma}$ were –.632 and –.688, respectively. When the same samples were shifted to have zero sample mean, these correlations changed little. The sample mean had a much smaller effect. Thus we see that any benefits to be gained from minimizing *ISE* rather than *MISE* are swamped by the much larger asymptotic error of the algorithms which pursue the former goal.

We also find that *CV* performance depends strongly on sample size and the underlying density. Specifically, the conditional probability that the *CV* smoothing parameter is "acceptable" given n increases rather rapidly from 10% to 90%; however, the location of this transition region may begin with surprisingly large sample sizes. Further work characterizing this transition would be interesting. With finite samples we are limited in our ability to adequately estimate all densities. Such fears are justified but clearly we are in a stronger position than if we made a parametric choice. As in spectral analysis, we have a bandwidth that indicates an upper bound on the size of a feature that may be hidden.

Another point of question is whether to use L_2 error or L_1 error, which is defined by $E \int | \hat{f} - f |$. For most cases, this choice makes little difference between the optimal estimates for a particular density estimator; that is, the prescription and competing (possible) density estimates (for alternative h's) are the same for either error criterion. We postpone discussion of the question of prior assumptions on f until Section 9.

8. Some More Examples

Good and Gaskins (1980) present a large particle physics data set (the LRL data), which is interesting because it is prebinned with δ=10 *MeV*. The authors found thirteen bumps in a penalized likelihood estimate. The optimal bin width using either histogram criteria (1.2) or (1.3) gives h =10 *MeV* as optimal.

We also examined these data with a triweight ASH estimator. In this case m =2 for *UCV* and m =4 for *BCV* using the ASH and FP-ASH, respectively. The square roots of these estimates are shown in Figure 10. Although the 13 bumps found by Good and Gaskins are apparent in Figure 10c, it is interesting to speculate why certain small bumps are included and others excluded. It is appropriate to recall that an optimally smoothed density has a slightly noisy second derivative, as shown in equation (3.15) when p =2.

We have implemented a bivariate product kernel *BCV* algorithm. Details and an example with a data set (thought to have a bimodal density) of 320 males with heart disease are available from us. The bimodal feature was not revealed by a *BCV* estimate, in the same manner observed in the univariate mixture example in Section 6.2.

9. Discussion, Conclusions, and Other Work

9.1. Other Related Work

Kronmal and Tarter (1968) introduced the first unbiased *CV* algorithm for a Fourier series density estimator. The algorithm provided unbiased estimates of the change in *MISE* as additional Fourier coefficients were introduced. Hart (1985) has given an unbiased procedure for Davis's Fourier integral series estimator. Wahba (1977) had the first working *BCV* algorithm, which she called generalized *CV*. In her Fourier series estimator, the smoothing parameter is not the number of terms in the series (taken to be n /2) but a design parameter in a tapering window applied to the Fourier coefficients. The leading terms in the theoretical *MISE* depend only on the magnitude of the Fourier coefficients $|f_\nu|^2$. By substituting an unbiased sample estimator for $|f_\nu|^2$, she derived a biased cross-validation criterion. The (small) bias results from truncation of the series in ν, and Wahba (1981) showed the truncated terms contributed only $O(n^{-2m})$ towards the *MISE*. Wahba's and Kullback-Liebler methods were tested by Scott and Factor (1981). Our biased *CV* algorithm is essentially the analog of Wahba's procedure.

9.2. Partial Explanation for Improvement

The improvement of *BCV* over *UCV* should not be viewed as artificial. The nature of the improvement is most easily seen with the histogram, for which,

$$AMISE(h) = \frac{1}{nh} + \frac{1}{12}h^2R(f') . \tag{9.1}$$

Recall that $h^* = O(n^{-1/3})$ for the histogram. Following (1.2) and (1.3), consider a third estimate of (9.1):

$$e_2(h) = \frac{23}{24nh} + \frac{1}{48n^2h}\sum[\nu_h(k+1) - \nu_h(k-1)]^2 . \tag{9.2}$$

Now $e_2(h)$ is based on a central difference approximation to $R(f')$, which is numerically superior to the forward difference approximation leading to $e_1(h)$. It may be shown that the "vertical" variances of e_0, e_1, and e_2 are $2R(f)/n^2h + O(h^2/n)$, $R(f)/12n^2h$, and $R(f)/192n^2h$, respectively. Again the squared bias is of lower order $O(n^{-2})$. This is a remarkable decrease in the variances. But for finite samples, the use of higher order derivative approximations will incur large bias and hence the gains are not realized except for extremely large samples. This is similar to the choice of p in equation (2.4). Theory suggests choosing p as large as possible, whereas in practice $p=2$ or 3 is a wiser choice. The higher order terms cannot in general be neglected. But for moderate samples $e_1(h)$ does represent a substantial improvement over $e_0(h)$, whereas $e_2(h)$ may not.

9.3. Discussion

We have attempted to evaluate the small-sample properties and reliability of two cross-validation algorithms. No currently available algorithm is highly reliable for very small samples. In this situation *BCV* always oversmoothes while *UCV* has very large variance. However, for "large" samples cross-validation is highly reliable with respect to *MISE*. Reliability with "medium" samples is often achieved with densities that are not too rough. From Tables II and III we see that our definition of a highly reliable CV algorithm is satisfied by the BCV estimates for sample sizes beyond 500-1000 except for the lognormal density, which requires several thousand points. The goal of finding $\tilde{h}_{ISE}$ with CV algorithms remains largely unsolved, as pointed out in Section 6.1. It is not at all clear that using $\tilde{h}_{ISE}$ is to be preferred to h_{MISE}, given the rather peculiar manner by which the integrated squared error is further reduced, as discussed in Section 7.

While biased CV has potentially greater reliability compared to unbiased CV, it comes at the cost of additional assumptions on f. However, the very general optimal consistency of unbiased CV comes at a surprisingly high cost in sample size requirements if f is smooth. Asymptotically, about $(4.98)^{10/3}$ or about 211 times more points are required so that $\sigma(\tilde{h}_{UCV})$ equals $\sigma(\tilde{h}_{BCV})$ for the triweight kernel. Thus we have a tension between "customized" and "generic" CV. It would be interesting to investigate how much unbiased CV can be improved perhaps, for example, by leaving more than one point out.

Perhaps most useful is to observe the divergence in behavior of UCV and BCV algorithms. Agreement or disagreement of the 2 CV parameters provides possible auxiliary information about any unusual features in the underlying density. Biased CV is essentially using the data to estimate the bias. This is (and should be) a difficult task because the relative contribution of the bias and variance towards the $MISE$ is in a ratio of 1:4 near optimal smoothing. Unbiased CV provides superior bias estimates but at the cost of increased variance. Given the importance of variance at $h = h_{MISE}$, it is important to control "vertical" variance more than current UCV algorithms do. Simple local averaging of the UCV curve is not the solution as one might have guessed from Figure 1.

We observe that the BCV procedure may be used to obtain approximate confidence intervals for both $\tilde{h}_{UCV}$ and $\tilde{h}_{BCV}$, assuming the latter is asymptotically normal. BCV provides consistent estimates of $R(f'')$ as well as $R(f)$, which may be used in (4.9) and (4.15). In fact, Theorem 3.2 follows from the fact that $\hat{R}(f'') = AN\{R(f''), 2R(\phi)R(f)/(n^2h^9)\}$. Some idea of the reliability of the CV smoothing parameter can be drawn from these estimates.

For sufficiently large data sets and reasonable densities, reliability is achievable. We wish to emphasize that excellent density estimates are still possible with smaller samples, but cannot be reliably calibrated by present methodology. We believe superior unbiased and biased CV kernel estimators can be found, since neither development attempted to optimize reliability. Perhaps the more computationally intensive bootstrap methods can be used to improve reliability for small samples.

Finally we remark that there are many other nonparametric applications where cross-validation is desirable, such as nonparametric regression, discrimination, hazard analysis and spectral analysis. It would be interesting to see how biased and unbiased cross-validation algorithms compare in these settings.

10. Proofs of Results

10.1. Proofs of Theorems 3.1 and 3.3 in Sections 3.1 and 3.3.1

We assume that conditions $(C1)$ and $(C2a)$ are satisfied. Occasionally in the proofs we tacitly assume the existence of higher order derivatives in f when we wish to investigate explicitly error terms; however these derivatives are not required.

10.1.1. Expectation of the Unbiased Cross-Validation Function

Recall the definitions of γ_+ and γ_- in equation (4.1). Since K is symmetric, it is easy to show that for $c \geq 0$, $\gamma_-(-c) = \gamma_+(c)$, that is, γ is symmetric. Now

$$
\begin{aligned}
E\,\gamma(c_{ij}) &= \int_{-\infty}^{\infty} f(x) \left[\int_{x-2h}^{x} \gamma_+\left(\frac{x-y}{h}\right) f(y)\,dy + \int_{x}^{x+2h} \gamma_-\left(\frac{x-y}{h}\right) f(y)\,dy \right] dx \\
&= h \int_x f(x) \left[\int_0^2 \gamma_+(c)[f(x-hc) + f(x+hc)]\,dc \right] dx \\
&= 2h \int_0^2 \gamma_+(c) \left[\sum_{k=0}^{3} \frac{(-1)^k}{(2k)!} c^k h^k R(f^{(k)}) + o(h^6) \right] dc \;, \qquad (10.1)
\end{aligned}
$$

by two change of variables, the symmetry of γ, a Taylor's series of $f\,(x \pm hc\,)$, and integrating by parts (e.g. $\int f f'' = -R\,(f'))$. It is not hard to show that

$$\int_0^2 c^k\, \gamma_+(c\,)dc \;=\; \frac{1}{2}\int_{-1}^{1} K\,(w\,)\int_{-1}^{1}(s-w\,)^k\, K\,(s\,)ds dw \;-\; \mu_k \;, \tag{10.2}$$

which equals $-1/2$, 0, $3\mu_2^2$, and $15\mu_2\mu_4$ for k =0,2,4,6, respectively. Hence,

$$E\,\gamma(c_{ij}\,) = -hR\,(f\,) + \frac{1}{4}\mu_2^2 h^5 R\,(f'') - 24\mu_2\mu_4 h^7 R\,(f''') + o\,(h^7)\,. \tag{10.3}$$

Thus (3.12) follows from (10.3), (3.11), and (2.3).

10.1.2. Variance of the Unbiased Cross-Validation Function

Next we find the variance of (3.11). The analysis of $E\,\gamma(c_{ij}\,)^2$ is parallel to (10.1) with $\gamma_+(c\,)^2$ rather than $\gamma_+(c\,)$. Hence $E\,\gamma(c_{ij}\,)^2 = \; hR\,(\gamma)R\,(f\,) + \; O\,(h^3)$; together with (10.3) we have

$$Var\;\; \gamma(c_{ij}\,) = hR\,(\gamma)R\,(f\,) - h^2 R\,(f\,)^2 + O\,(h^3)\,. \tag{10.4}$$

For simplicity of notation, let $\gamma_{ij} \equiv \gamma(c_{ij}\,)$. Now $Cov\,(\gamma_{ij}\,,\gamma_{kl}\,) = 0$; here (and from now on) we assume distinct letters represent unequal subscripts. Let

$$I_1 \equiv \int f^{\,iv}(x\,)f\,(x\,)^2 dx\,; \quad I_2 \equiv R\,(f\,)R\,(f''); \quad I_3 \equiv R\,(f^{\,3/2}) - R\,(f\,)^2\,.$$

Consider

$$E\,\gamma_{ij}\,\gamma_{ik} \;=\; \int f\,(x\,)\left[\int \gamma(\frac{x-y}{h})f\,(y\,)dy\right]\left[\int \gamma(\frac{x-z}{h})f\,(z\,)dz\right]dx$$

$$= \int f\,(x\,)\left[h\int_0^2 \gamma_+(c\,)[f\,(x-hc\,) + f\,(x+hc\,)]\,dc\right]^2 dx \tag{10.5}$$

$$= h^2 \int f\,(x\,)^3 dx \;-\; \frac{1}{2}\mu_2^2 I_1 + O\,(h^8)\,. \tag{10.6}$$

With (10.3) we have

$$Cov\,(\gamma_{ij}\,,\gamma_{ik}\,) = h^2 I_3 - \frac{1}{2}\mu_2^2 h^6 [I_1 - I_2] + O\,(h^8). \tag{10.7}$$

Quantities such as $Cov\,(\gamma_{ij}\,,\gamma_{ki}\,)$ also equal (10.7). Now it is well-known that

$$Var\,[\sum_{i<j}\sum \gamma_{ij}\,] = \frac{1}{2}n\,(n-1)\,Var\;\; \gamma_{ij} \;+ n\,(n-1)(n-2)Cov\,(\gamma_{ij}\,,\gamma_{ik}\,)\,. \tag{10.8}$$

With (10.4), (10.7), (10.8), and (3.11), we have

$$Var \ \ UCV(h) = \frac{4}{n} I_3 + \frac{2R(\gamma)R(f)}{n^2 h} + \frac{2\mu_2^2 h^4}{n}[I_2 - I_1] + o(n^{-9/5}), \tag{10.9}$$

which explicitly gives the remainder terms in (3.13), and proving Theorem 3.1.

10.1.3. Variance of Augmented *UCV* Criterion

Comparing (3.11) and (3.22), we see that

$$Var \ \ AUCV(h) = Var \ \ UCV(h) + \frac{4}{n^2} Var \sum_{i=1}^{n} f(x_i) + \frac{8}{n^3 h} Cov\left(\sum_{i<j}\sum \gamma_{ij}, \sum_i f(x_i)\right). \tag{10.10}$$

Since $Ef(x_i)^k = \int f(x)^{k+1} dx$, we have

$$\frac{4}{n^2} Var \sum_{i=1}^{n} f(x_i) = \frac{4}{n}[R(f^{3/2}) - R(f)^2] = \frac{4}{n} I_3 . \tag{10.11}$$

In (10.10), $Cov(\gamma_{ij}, f(x_k)) = 0$. Consider the $n(n-1)$ terms for which $k = i$ or $k = j$:

$$E\gamma_{ij} f(x_i) = 2h\int_0^2 \gamma_+(c)\left[\int_{-\infty}^{\infty} [f(x) + \frac{1}{2}h^2 c^2 f''(x) + \frac{1}{24}h^4 c^4 f^{iv}(x) + ..] f(x)^2 dx\right] dc$$

$$= -hR(f^{3/2}) + \frac{1}{4} h^5 \mu_2^2 I_1 + O(h^7). \tag{10.12}$$

Since $Ef(x_i) = R(f)$, combining (10.12) and (10.3),

$$Cov(\gamma_{ij}, f(x_i)) = -hI_3 + \frac{1}{4}\mu_2^2 h^5 [I_1 - I_2].$$

Together with (10.10), (10.9), and (10.11), we have proven Theorem 3.3.

10.2. Proof of Lemma 3.2 in Section 3.2

Clearly

$$\hat{f}^{(p)}(x) = \frac{1}{nh^{p+1}} \sum_{i=1}^{n} K^{(p)}\left(\frac{x - x_i}{h}\right).$$

$$E \ R(\hat{f}^{(p)}) = \frac{1}{nh^{2p+2}} \int\int K^{(p)}\left(\frac{x-y}{h}\right)^2 f(y) dy dx \ +$$

$$\frac{n(n-1)}{n^2 h^{2p+2}} \int\int\int K^{(p)}\left(\frac{x-y}{h}\right) K^{(p)}\left(\frac{x-z}{h}\right) f(z) f(y) \ dz \ dy \ dx$$

$$= \frac{R(K^{(p)})}{nh^{2p+1}} + \frac{n-1}{nh^{2p}} \int \left[\int K^{(p)}(w) f(x - hw) dw\right]^2 dx$$

after a change of variables. Now the bracketed term may be approximated by

$$\sum_{i=o}^{p+2} \frac{1}{i\,!}(-h)^i f^{(i)}(x)\int w^i K^{(p)}(w)dw + o(h^{p+2}) .$$

Now $\int w^i K^{(p)}(w)dw = 0$ if $i < p$ or for i odd, and it equals $(-1)^p p\,!$ for $i = p$ and $(-1)^{p+2}(p+2)!\mu_2/2$ for $i = p+2$. Hence the sum collapses to $h^p f^{(p)}(x) + O(h^{p+2})$. Squaring, integrating, and noting that $(n-1)/n = 1 + O(n^{-1})$ completes the proof. We remark that since $\mu_k = 0$ for $0 < k < p$, the error is actually $O(h^p)$ if $f^{(2p)}$ exists.

10.3. Proof of Theorem 3.2 in Section 3.2

The analysis of the moments of the biased cross-validation function is similar to that in Section 10.1, although much easier since $BCV(h)$ involves fewer terms and because more "moments" of (10.13) below vanish. We assume conditions $(C1)$ and $(C2a,b)$ are satisfied. We remark that condition $(C2b)$ is necessary for Theorem 4.2 but stronger than necessary by one order of derivative for Theorem 3.2. From (3.16) define

$$\phi_+(c) = \int_{-1}^{1-c} K''(w)K''(w+c)dw \qquad 0 \le c \le 2 \tag{10.13}$$

and $\phi_-(c)$ for $-2 \le c \le 0$. Again ϕ is symmetric. Now

$$\int_0^2 c^k \phi_+(c)dc = \int_{-1}^{1} K''(w)\int_0^{1-w} c^k K''(w+c)dc\,dw . \tag{10.14}$$

For $k=0$, observe $\int_0^{1-w} K''(w+c)dc = -K'(w)$ and $-\int_{-1}^{1} K'(w)K''(w)dw = 0$ since $K'(\pm 1) = 0$. For $k \ge 2$,

$$\int_0^{1-w} c^k K''(w+c)dc = k(k-1)\int_0^{1-w} c^{k-2}K(w+c)dc .$$

Noting (for even m)

$$\int_{-1}^{1} K(s)\int_{-1}^{s} w^m K(w)dw = \frac{1}{2}\int_{-1}^{1} K(s)ds\int_{-1}^{1} w^m K(w)dw = \frac{\mu_m}{2} ,$$

and integrating by parts, we see that

$$\int_0^2 c^k \phi_+(c)dc = 0,\ 0,\ 12,\ 360\mu_2 \quad \text{for } k = 0,\ 2,\ 4,\ 6, \quad \textit{respectively.} \tag{10.15}$$

The analysis proceeds exactly as in Section 10.1 with ϕ_+ replacing γ_+. Let $\phi_{ij} \equiv \phi(c_{ij})$. From (10.1), it follows that

$$E\,\phi_{ij} = h^5 R\,(f'') - h^7 \mu_2 R\,(f''') + o\,(h^7)\,, \tag{10.16}$$

from which (3.18) follows directly. From (10.4) and (10.16), it follows that

$$Var\ \ \phi_{ij} = hR\,(\phi) R\,(f\ ) + O\,(h^3)\,. \tag{10.17}$$

Corresponding to (10.5) we have

$$E\,\phi_{ij}\,\phi_{ik} = h^2 \int f\ (x\,) \left[2\int_0^2 \phi_+ [f\ (x\,) + \frac{1}{2} h^2 c^2 f''(x\,) + \frac{1}{24} h^4 c^4 f^{\,iv} (x\,) + ..]\,dc \right]^2 dx$$

$$= h^{10} \int_{-\infty}^{\infty} f^{\,iv}\,(x\,)^2 f\ (x\,)\,dx\ +\ o\,(h^{10})\,. \tag{10.18}$$

Therefore $Cov\,(\phi_{ij}\,,\phi_{ik}\,) = O\,(h^{10})$. Following (10.8),

$$Var\,[\sum_{i<j}\sum \phi(c_{ij}\,)] = \frac{1}{2} n^2 hR\,(\phi) R\,(f\ ) + O\,(n^2 h^3) + O\,(n^3 h^{10})\,,$$

which, together with (3.17), proves equation (3.19).

The asymptotic normality follows from Theorem 3.1 of Hall (1984) for AN of degenerate U-statistics. Let $\mu(t\,) = EK''((t - X\,)/h\,)$, where $h = cn^{-1/5}$. Then $h\,\phi(c_{ij}\,)$ may be decomposed into

$$h\,\phi(\frac{x_i - x_j}{h}) = \int [K''(\frac{t - x_i}{h}) - \mu(t\,)][K''(\frac{t - x_j}{h}) - \mu(t\,)]\,dt$$

$$+ \int \mu(t\,)[K''(\frac{t - x_i}{h}) + K''(\frac{t - x_j}{h})]\,dt\ -\ \int \mu(t\,)^2 dt\ .$$

The mean comes from the last two integrals and is easily checked to be $h^6 R\,(f'') + o\,(h^6)$. The variance comes from the first integral, which we denote by $H_n\,(x_i\,,x_j\,)$. Now it may be verified that the random variable $E\,[H_n\,(X\,,Y\,)\mid X\,] \equiv 0$, so that H_n is a degenerate Martingale. Using the notation of Hall's (1984) equation (2.1), calculations similar to the above give $EH_n^{\,2} = h^3 R\,(\phi) R\,(f\ )$, $EH_n^{\,4} = h^5 R\,(\phi^2) R\,(f\ )$, and $EG_n\ = O\,(h^7)$; therefore, the conditions for Hall's Theorem 3.1 hold and $BCV(h\,)$ is AN.

10.4. Proof of Corollaries 3.2 and 3.3 in Section 3.2 and 3.3.1

From (2.3) and Theorem 3.2 we have that

$$\begin{aligned}
\underset{n \to \infty}{plim}\ [BCV(ch^*)/MISE(ch^*)] &= 1 \\
\underset{n \to \infty}{plim}\ [AMISE(ch^*)/MISE(ch^*)] &= 1 \qquad \textbf{(10.19)} \\
AMISE(ch^*)/AMISE(h^*) &= \frac{c^5+4}{5c}
\end{aligned}$$

so that $MISE(ch^*) > MISE(h^*)$ for $c \neq 1$ and large n. Suppose $\tilde{h}_{BCV}/h^*$ does not converge to one. Then $Prob\,\{BCV(\tilde{h}_{BCV}) < BCV(h^*)\} \to 1$ as $n \to \infty$, which contradicts the consistency results in (10.19).

The proof of Corollary 3.3 was first given by Hall (1983).

10.5. Proof of Lemma 4.1 in Section 4.1

As before, define $\rho_+(c)$ from (4.2) when $0 \leq c \leq 2$. Then it may be shown that

$$\int_0^2 c^k \rho_+(c)\,dc = \frac{1}{2},\ 0,\ -15\mu_2^2,\ -105\mu_2\mu_4 \quad \text{for} \quad k=0,2,4,6, \ \textit{respectively.}$$

Let $\rho_{ij} \equiv \rho(c_{ij})$. Omitting details,

$$E\,\rho_{ij} = hR(f) - \frac{5}{4}\mu_2^2 h^5 R(f'') + \frac{7}{24}\mu_2\mu_4 h^7 R(f''') + o(h^7)$$

from which the expectation in (4.4) may be computed. Then

$$\begin{aligned}
Var\,\rho_{ij} &= hR(\rho)R(f) - h^2R(f)^2 + O(h^3) \\
Cov(\rho_{ij},\rho_{ik}) &= h^2 I_3 + \frac{5}{2}\mu_2^2 h^6[I_2 - I_1] + o(h^6) \\
Cov(\gamma_{ij},\rho_{ik}) &= -h^2 I_3 + \frac{3}{2}\mu_2^2 h^6[I_1 - I_2] + o(h^6) \\
Cov(\gamma_{ij},\rho_{ij}) &= hR(\sqrt{\gamma\rho})R(f) + h^2R(f)^2 + O(h^3).
\end{aligned}$$

Now the variance of the left hand side of (4.4) may be expressed as

$$\frac{1}{2}n(n-1)[Var\,\gamma_{ij} + Var\,\rho_{ij}] + n(n-1)(n-2)[Cov(\gamma_{ij},\gamma_{ik}) + Cov(\rho_{ij},\rho_{ik})] +$$

$$n(n-1)Cov(\gamma_{ij},\rho_{ij}) + 2n(n-1)(n-2)Cov(\gamma_{ij},\rho_{ik})\,. \qquad (10.20)$$

Evaluating (10.20), we find

$$Var \sum_{i<j}\sum [\gamma(c_{ij}) + \rho(c_{ij})] = \frac{1}{2}n^2h\,[R(\gamma) + R(\rho) + 2R(\sqrt{\gamma\rho})]R(f)$$

where the bracketed term may be written as $R(\gamma+\rho)$. But $\frac{d}{dc}\gamma_+(c) = \rho_+(c)/c$. Hence

$$\int_0^2 \gamma_+(c)\rho_+(c)dc = \int_0^2 c\,\gamma_+(c)\gamma_+'(c)dc = -\frac{1}{2}\int_0^2 \gamma_+(c)^2 dc$$

since $\gamma_+(2)=0$. Since γ is symmetric, it follows that $R(\gamma+\rho) = R(\rho)$, which completes the argument.

10.6. Proof of Lemma 4.2 in Section 4.2

Briefly,

$$\int_0^2 c^k \psi_+(c)dc = 0,\ 0,\ -60,\ -2520\mu_2 \quad \text{for} \quad k=0,2,4,6, \qquad \textit{respectively.}$$

$$\begin{aligned} E\,\psi(c_{ij}) &= -5h^5 R(f'') + O(h^7) \\ Var\,\psi(c_{ij}) &= hR(\psi)R(f) + O(h^3) \\ Cov(\psi_{ij},\psi_{ik}) &= O(h^{10})\,; \quad Cov(\phi_{ij},\psi_{ik}) = O(h^{10}) \\ Cov(\phi_{ij},\psi_{ij}) &= hR(\sqrt{\phi\psi})R(f) + O(h^3) \end{aligned}$$

and $R(\phi+\psi) = R(\psi)$ as above. The lemma follows directly.

11. Acknowledgments: This research was partially supported by ONR contract N00014-85-K-0100 and ARO contract DAAG-29-85-K0212. This paper was written while the first author was visiting Stanford, and we wish to acknowledge their generous support, particularly the partial support of J. Friedman under ONR contract N00014-83-K-0472. We also benefited from several discussions with Steve Marron.

12. References

Bowman, A. W. (1985), "A Comparative Study of Some Kernel-Based Nonparametric Density Estimates," Journal of Statistical Computation and Simulation, 21, 313-327.

Bowman, A. W. (1984), "An Alternative Method of Cross-Validation for the Smoothing of Density Estimates," Biometrika 65, 521-528.

Carroll, R.J., and Ruppert, D. (1985), "Transformations in Regression: A Robust Analysis," Technometrics, 27, 1-12.

Fryer, M.J. (1976), "Some Errors Associated with the Non-parametric Estimation of Density Functions," Journal of the Institute of Mathematics and Applications, 18, 371-380.

Gale, W.A. and Oregibon, D. (1983), "An Expert System for Regression Analysis," in Computer Science and Statistics: Proceedings of the 14th Symposium on the Interface, Heiner, et al., eds., Springer-Verlag, New York, 110-117.

Good, I. J., and Gaskins, R. A. (1980), "Density Estimation and Bump-Hunting by the Penalized Likelihood Method Exemplified by Scattering and Meteorite Data," Journal of the American Statistical Association 75, 42-56.

Hall, P. (1983), "Large Sample Optimality of Least Squares Cross-Validation in Density Estimation," Annals of Statistics 11, 1156-1174.

Hall, P. (1984), "Central Limit Theorem for Integrated Square Error of Multivariate Nonparametric Density Estimators," Journal of Multivariate Analysis 14, 1-16.

Hall, P., and Marron, J. S. (1985), "The Amount of Noise Inherent in Bandwidth Selection for a Kernel Density Estimator," University of North Carolina Technica Report No. 100.

Härdle, W., and Marron, J. S. (1985), "Optimal Bandwidth Selection in Nonparametric Regression Function Estimation," Annals of Statistics 12,

Hart, J.D. (1985), "Data-Based Choice of the Smoothing Parameter for a Kernel Density Estimator," Australian Journal of Statistics 27, 53-59.

Kronmal, R., and Tarter, M.E. (1968), "The Estimation of Probability Densities and Cumulatives by Fourier Series Methods," Journal of the American Statistical Society 63, 925-952.

Rice, J. (1984), "Bandwidth Choice for Nonparametric Regression," Annals of Statistics 12, 1215-1230.

Rudemo, M. (1982), "Empirical Choice of Histogram and Kernel Density Estimators," Scandinavian Journal of Statistics 9, 65-78.

Scott, D. W. (1979), "On Optimal and Data-Based Histograms," Biometrika 66, 605-610.

Scott, D.W. (1985), "Averaged Shifted Histograms: Effective Nonparametric Estimators in Several Dimensions," *The Annals of Statistics* 13:1024-1040.

Scott, D.W. (1986), "Choosing Smoothing Parameters for Density Estimators," To appear in *Proceedings of the 17th Symposium on the Interface of Computer Science and Statistics*, North-Holland.

Scott, D. W., and Factor, L. E. (1981), "Monte Carlo Study of Three Data- Based Nonparametric Probability Density Estimators," Journal of the American Statistical Association 76,9-15.

Scott, D. W., Tapia, R. A., and Thompson, J. R. (1977), "Kernel Density Estimation Revisited," Nonlinear Analysis, Theory, Methods and Applications 1, 339-372.

Serfling, R.J. (1980), *Approximation Theorems of Mathematical Statistics*, John Wiley Sons, New York.

Silverman, B. W. (1978), "Choosing the Window Width When Estimating a Density," Biometrika 65, 1-11.

Silverman, B.W. (1982), "Kernel density estimation using the fast Fourier transform," Statistical Algorithm AS 176, Applied Statistics, 31, 93-97.

Stone, C. J. (1984), "An Asymptotically Optimal Window Selection Rule for Kernel Density Estimates," Annals of Statistics 12, 1285-1297.

Stone, M. (1974), "Cross-Validation and Multinomial Prediction," Biometrika 61:509-515.

Titterington, D. M. (1985), "Common Structure of Smoothing Techniques in Statistics," International Statistical Review 53:141-170.

Wahba, G. (1977), "Optimal Smoothing of Density Estimates," in Classification and Clustering, ed. J. Van Ryzin, New York: Academic Press.

Wahba, G. (1981), "Data-Based Optimal Smoothing of Orthogonal Series Density Estimates," Annals of Statistics 9, 146-156.

Woodroofe, M. (1970), "On Choosing a Delta Sequence," Annals of Mathematical Statistics 41, 1665-1671.

Legends for Tables and Figures

Table I. For several kernels, asymptotic ratios of "vertical" and "horizontal" standard deviation of unbiased and biased cross-validation estimates and smoothing parameters as given in expressions (3.25) and (4.16).

Table II. Summary of a Monte Carlo experiment using a triweight kernel averaged shifted histogram estimator with standard Gaussian data. The sample means and variances of the cross-validation smoothing parameters are given, together with the theoretical standard deviations given in Theorems 4.1 and 4.2. The theoretical predictions of the standard deviations of $\tilde{h}_{BCV}$ and $\tilde{h}_{UCV}$ are denoted by σ_{BCV} and σ_{UCV}, while sample versions are indicated by a circumflex.

Table III. Summary of partial results of a Monte Carlo experiment for three sampling densities. Other details are the same as in Table II.

Figure 1. Biased *(a)* and unbiased *(b)* cross-validation curves for a histogram estimator of 10,000 $N(5,1)$ points. The vertical lines in the bottom right hand corner of the figures indicate theoretical standard deviations computed from Theorems 3.1, 3.2, and 3.3 as discussed in the text. The predicted optimal *MISE* smoothing parameter is indicated by a star.

Figure 2. Examples of biased and unbiased cross-validation curves (on a $\log_{10}$ scale) for a Gaussian kernel estimator of 25 $N(0,1)$ points in (a) and 400 points in (b). The exact mean integrated squared error is shown by the dotted line. The corresponding cross-validation density estimates are shown in (c), along with the true density (dotted line).

Figure 3. Histograms of biased and unbiased cross-validation smoothing parameters for $N(0,1)$ samples of several sizes using an ASH triweight kernel estimator. The *BCV* histogram is in the positive direction while the *UCV* histogram is in the negative direction. The location of h_{MISE} is indicated by a star on the horizontal axis. These figures are discussed more fully in Section 6.1.

Figure 4. Scatter plots of the various smoothing parameters are shown for the same Monte Carlo data as in Figure 3 with sample sizes $n=400$ and $n=25600$.

Figure 5. Using the same samples as in Figure 4, scatter plots of the $\log_{10}(ISE)$ corresponding to the various smoothing parameters are shown. The first row is for the $n=400$ samples and the second row is for the $n=25600$ samples. The diagonal line indicates $y=x$.

Figure 6. Similar to Figure 3, except with n=1600 from Cauchy, Lognormal, and Mixture Densities.

Figure 7. Biased and unbiased cross-validation density estimates of 1600 points from a Lognormal distribution.

Figure 8. Biased and unbiased cross-validation density estimates of 400 points from a Mixture distribution.

Figure 9. For the samples in Figure 4, scatter plots of $\tilde{h}_{ISE}$ and the sample standard deviation for each sample.

Figure 10. On a square root scale, triweight ASH estimates of the LRL data, with $m=1$, 2, and 4. The bumps found by Good and Gaskins with a penalized-likelihood density estimator are indicated by horizontal lines above the bump.

Table I. Asymptotic Ratio of "Vertical" Standard Deviations of UCV and BCV Estimators

$K(t)=a_m(1-t^2)^m$	$R(\gamma)^{1/2}$	$\mu_2^2R(\phi)^{1/2}/4$	*ratio* Eqn(3.25)	$R(\rho)^{1/2}$	$\mu_2^2R(\psi)^{1/2}/4$	*ratio* Eqn(4.16)
m =2 (biweight)	1.0033	.0827	12.13	1.2352	-	-
m =3 (triweight)	1.0737	.0921	11.65	1.2047	.2420	4.98
m =4	1.1337	.1013	11.20	1.2195	.2550	4.78
m =5	1.1859	.1092	10.86	1.2458	.2685	4.64
N(0,1) ("m =∞")	.6376	.0715	8.92	.6178	.1558	3.96

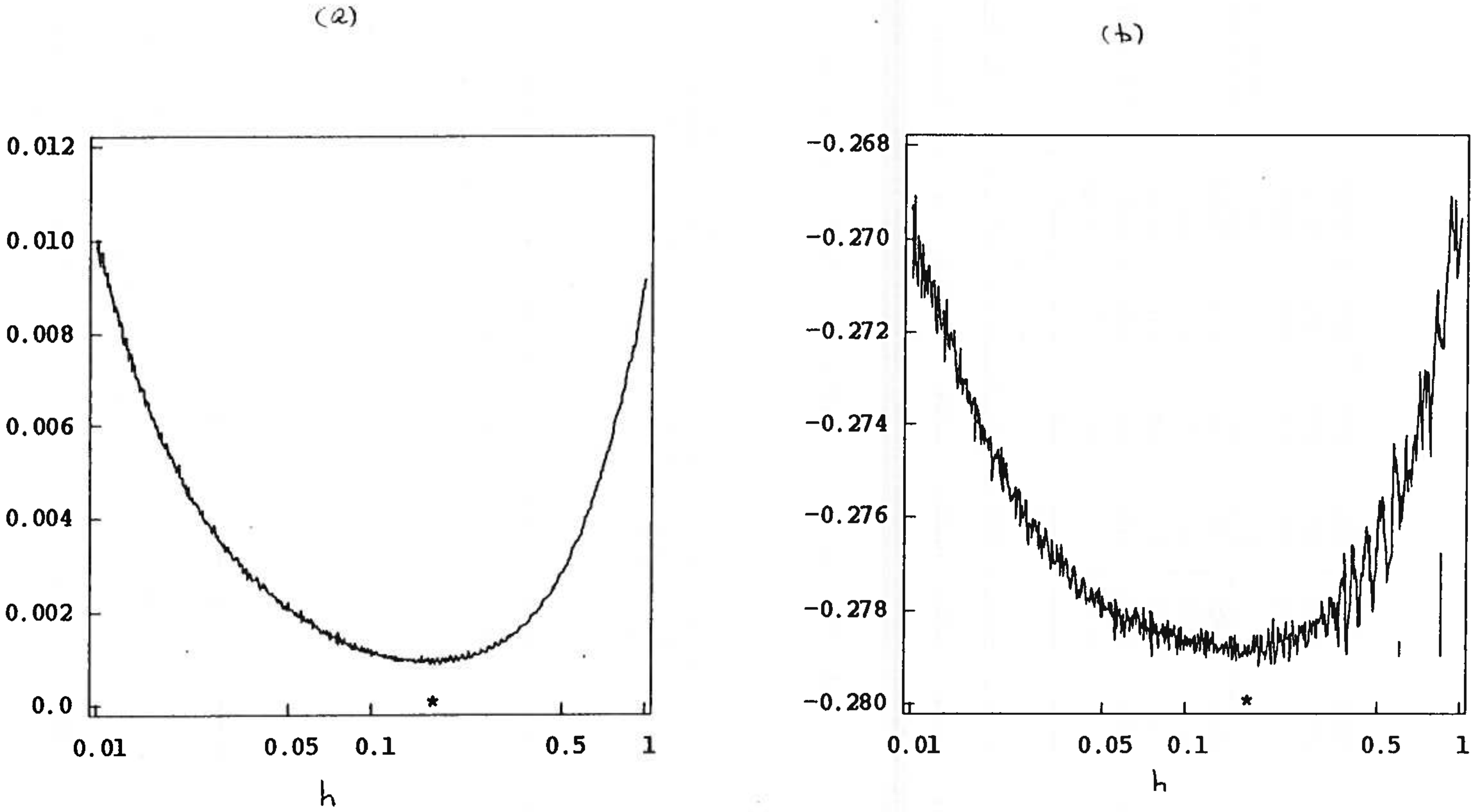
Table II. Monte Carlo Results or Triweight Kernel ASH Estimates of $N(0,1)$ Data

n	h_{MISE}	$\bar{h}_{BCV}$	$\bar{h}_{UCV}$	$\hat{\sigma}_{BCV}$	$\hat{\sigma}_{UCV}$	*ratio*	σ_{BCV}	σ_{UCV}
25	1.775	-	1.907	-	.6700	-	.0951	.4732
100	1.309	1.499	1.262	.1691	.4170	2.47	.0627	.3122
400	.976	1.041	.935	.0792	.2422	3.06	.0414	.2060
1600	.732	.753	.683	.0372	.1862	5.00	.0273	.1359
6400	.552	.561	.535	.0246	.1054	4.27	.0180	.0896
25600	.416	.419	.416	.0128	.0549	4.28	.0119	.0591

Table III. Partial Monte Carlo Results for Other Densities

density	n	h_{MISE}	$\bar{h}_{BCV}$	$\bar{h}_{UCV}$	$\hat{\sigma}_{BCV}$	$\hat{\sigma}_{UCV}$	*ratio*	σ_{BCV}	σ_{UCV}
Cauchy	400	1.012	1.230	1.056	.1292	.2538	1.96	.0300	.1492
	1600	.740	.815	.751	.0547	.1448	2.65	.0198	.0984
	6400	.549	.580	.551	.0263	.0862	3.28	.0130	.0649
	25600	.411	.418	.415	.0144	.0371	2.57	.0086	.0428
Lognormal	400	.324	.540	.326	.1052	.0776	0.74	.0050	.0248
	1600	.218	.302	.212	.0331	.0402	1.21	.0033	.0163
	6400	.151	.184	.150	.0137	.0209	1.52	.0022	.0108
	25600	.107	.121	.107	.0048	.0127	2.63	.0014	.0071
Mixture	400	.612	-	.618	-	.1512	-	.0167	.0830
	1600	.443	.504	.434	.0374	.0749	2.00	.0110	.0548
	6400	.327	.345	.320	.0155	.0425	2.75	.0073	.0361
	25600	.245	.252	.242	.0068	.0294	4.34	.0048	.0238

Figure 1


(a)
0.012
0.010
0.008
0.006
0.004
0.002
0.0
0.01
0.05
0.1
0.5
1
h
*
(b)
-0.268
-0.270
-0.272
-0.274
-0.276
-0.278
-0.280
0.01
0.05
0.1
0.5
1
h
*

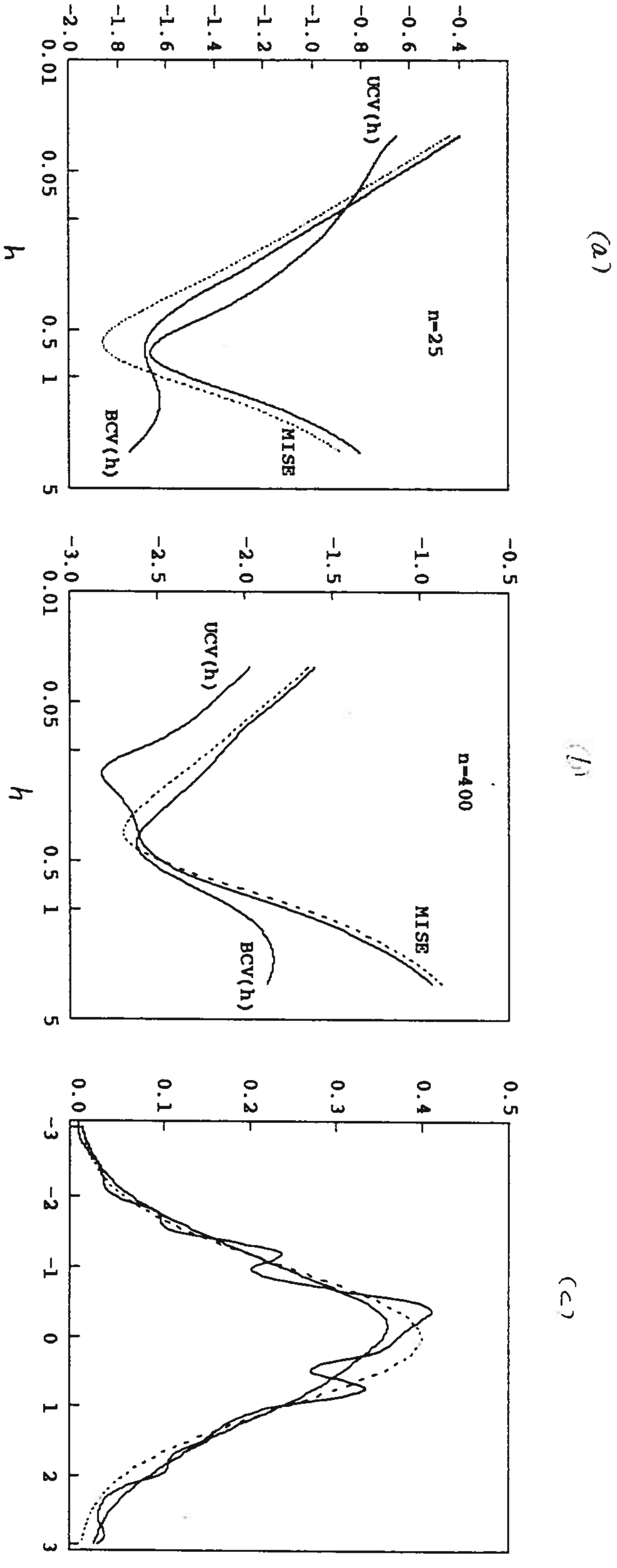

(a)
n=25
UCV(h)
MISE
BCV(h)
h
-0.4
-0.6
-0.8
-1.0
-1.2
-1.4
-1.6
-1.8
-2.0
0.01
0.05
0.5
1
5
(b)
n=400
UCV(h)
MISE
BCV(h)
h
-0.5
-1.0
-1.5
-2.0
-2.5
-3.0
0.01
0.05
0.5
1
5
(c)
0.5
0.4
0.3
0.2
0.1
0.0
-3
-2
-1
0
1
2
3

Figure 2

Figure 3

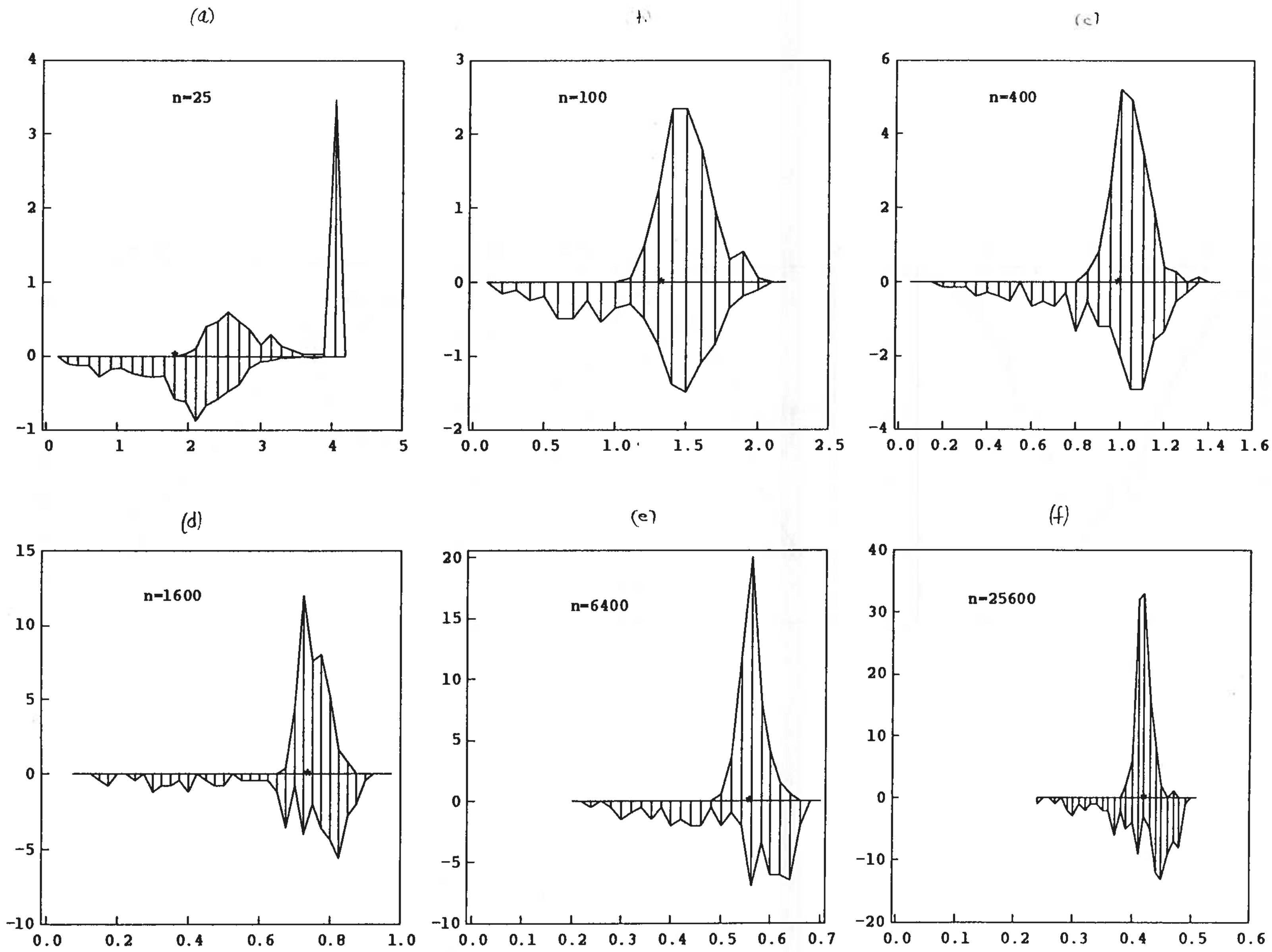

(a)
n=25
4
3
2
1
0
-1
0
1
2
3
4
5
(b)
n=100
3
2
1
0
-1
-2
0.0
0.5
1.0
1.5
2.0
2.5
(c)
n=400
6
4
2
0
-2
-4
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
(d)
n=1600
15
10
5
0
-5
-10
0.0
0.2
0.4
0.6
0.8
1.0
(e)
n=6400
20
15
10
5
0
-5
-10
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
(f)
n=25600
40
30
20
10
0
-10
-20
0.0
0.1
0.2
0.3
0.4
0.5
0.6

Figure [illegible]

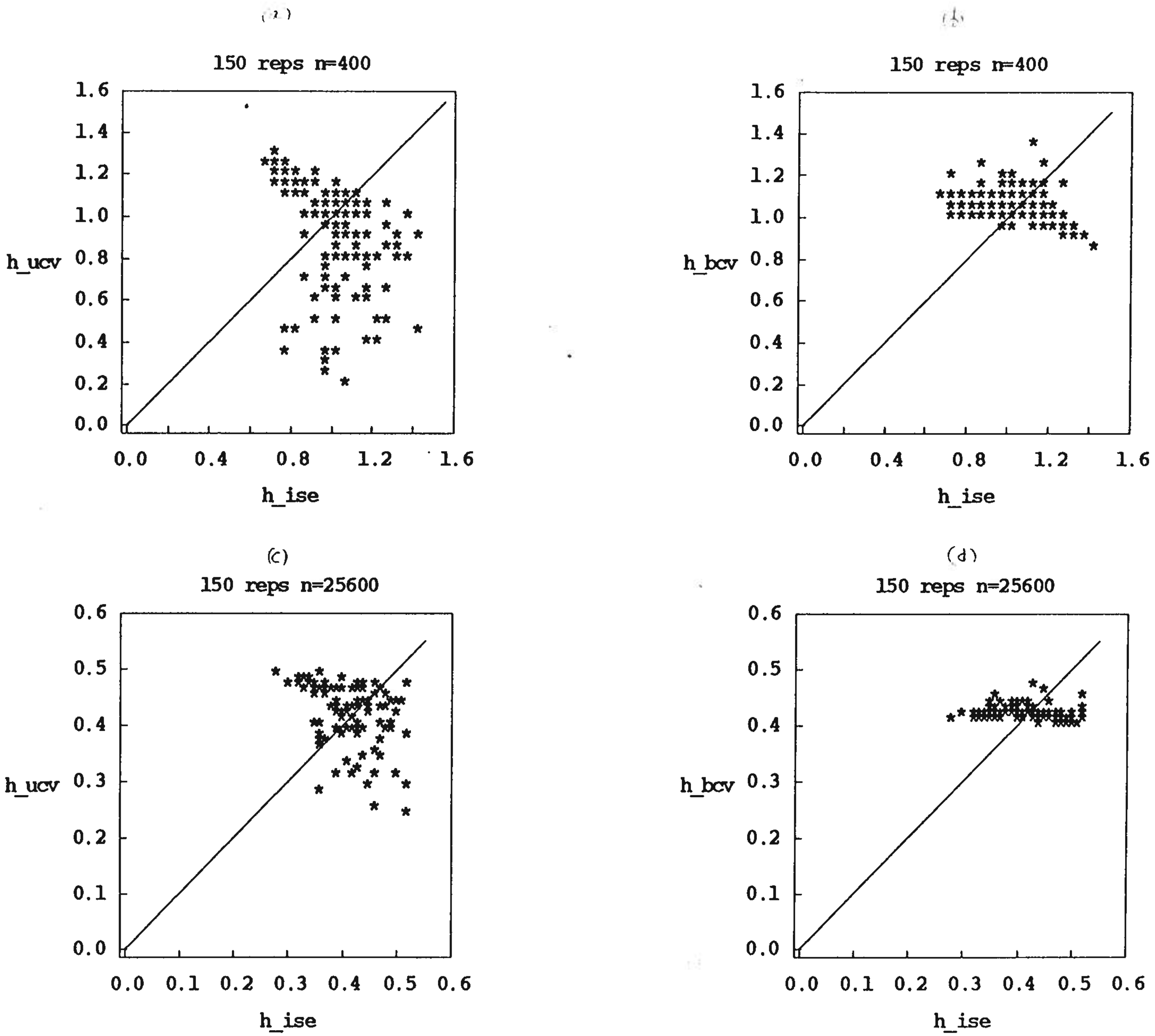

(a)
150 reps n=400
h_ucv
h_ise
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
(b)
150 reps n=400
h_bcv
h_ise
(c)
150 reps n=25600
h_ucv
h_ise
0.0
0.1
0.2
0.3
0.4
0.5
0.6
(d)
150 reps n=25600
h_bcv
h_ise

Figure 5

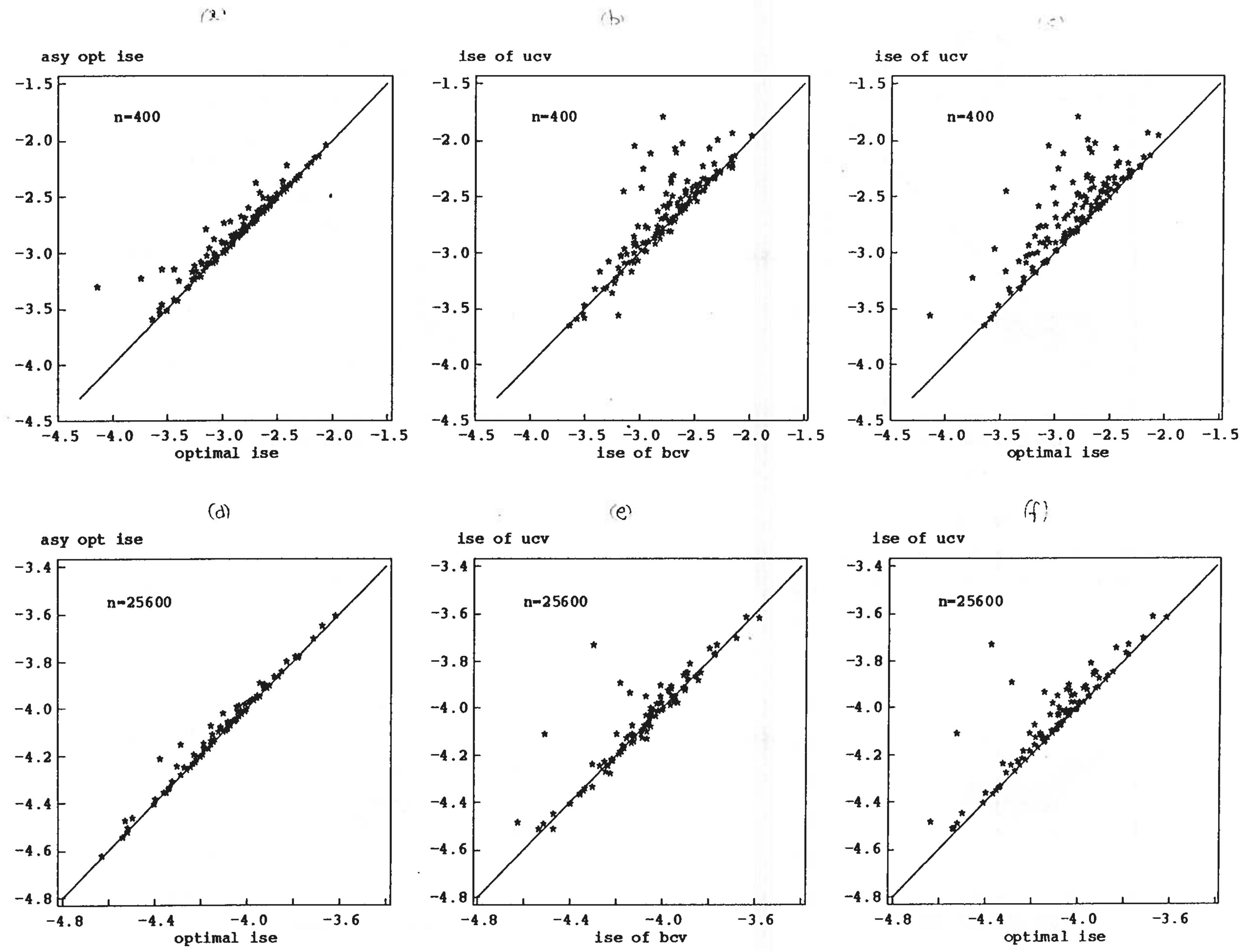

(a)
asy opt ise
n=400
-1.5
-2.0
-2.5
-3.0
-3.5
-4.0
-4.5
-4.5 -4.0 -3.5 -3.0 -2.5 -2.0 -1.5
optimal ise
(b)
ise of ucv
n=400
-1.5
-2.0
-2.5
-3.0
-3.5
-4.0
-4.5
-4.5 -4.0 -3.5 -3.0 -2.5 -2.0 -1.5
ise of bcv
(c)
ise of ucv
n=400
-1.5
-2.0
-2.5
-3.0
-3.5
-4.0
-4.5
-4.5 -4.0 -3.5 -3.0 -2.5 -2.0 -1.5
optimal ise
(d)
asy opt ise
n=25600
-3.4
-3.6
-3.8
-4.0
-4.2
-4.4
-4.6
-4.8
-4.8 -4.4 -4.0 -3.6
optimal ise
(e)
ise of ucv
n=25600
-3.4
-3.6
-3.8
-4.0
-4.2
-4.4
-4.6
-4.8
-4.8 -4.4 -4.0 -3.6
ise of bcv
(f)
ise of ucv
n=25600
-3.4
-3.6
-3.8
-4.0
-4.2
-4.4
-4.6
-4.8
-4.8 -4.4 -4.0 -3.6
optimal ise

Figure 6

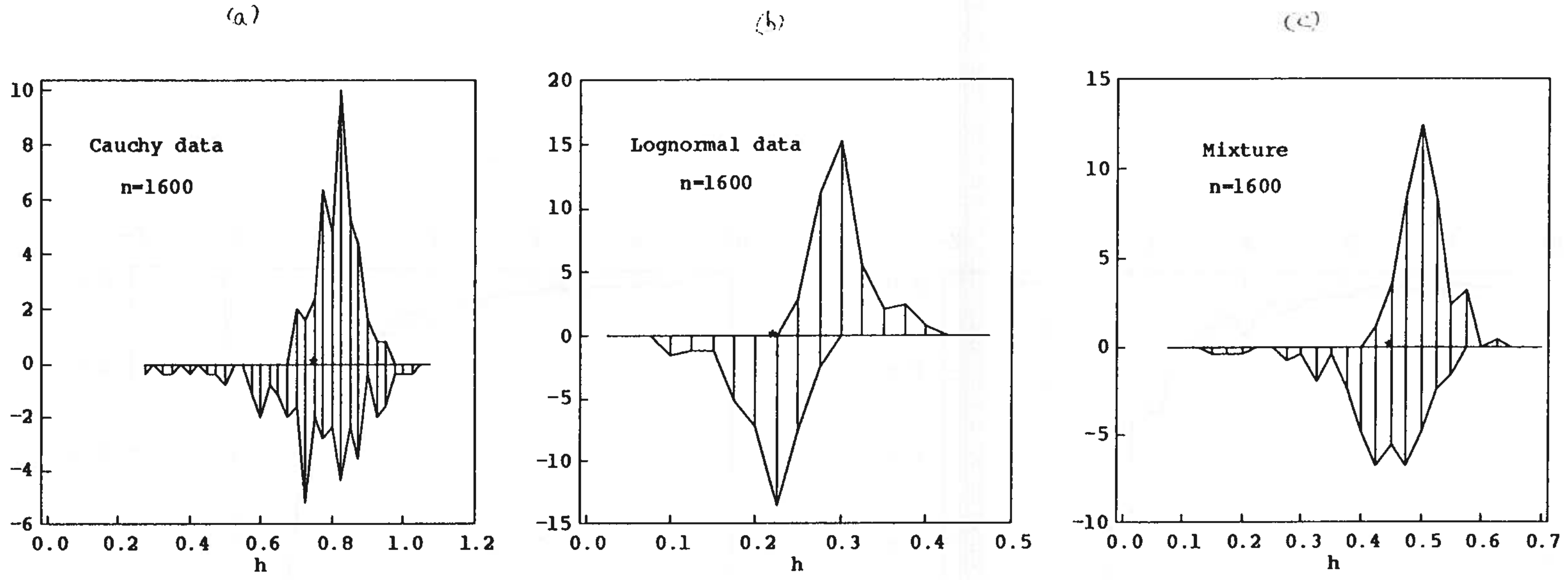

(a)
Cauchy data
n=1600
10
8
6
4
2
0
-2
-4
-6
0.0
0.2
0.4
0.6
0.8
1.0
1.2
h
(b)
Lognormal data
n=1600
20
15
10
5
0
-5
-10
-15
0.0
0.1
0.2
0.3
0.4
0.5
h
(c)
Mixture
n=1600
15
10
5
0
-5
-10
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
h

Figure 7

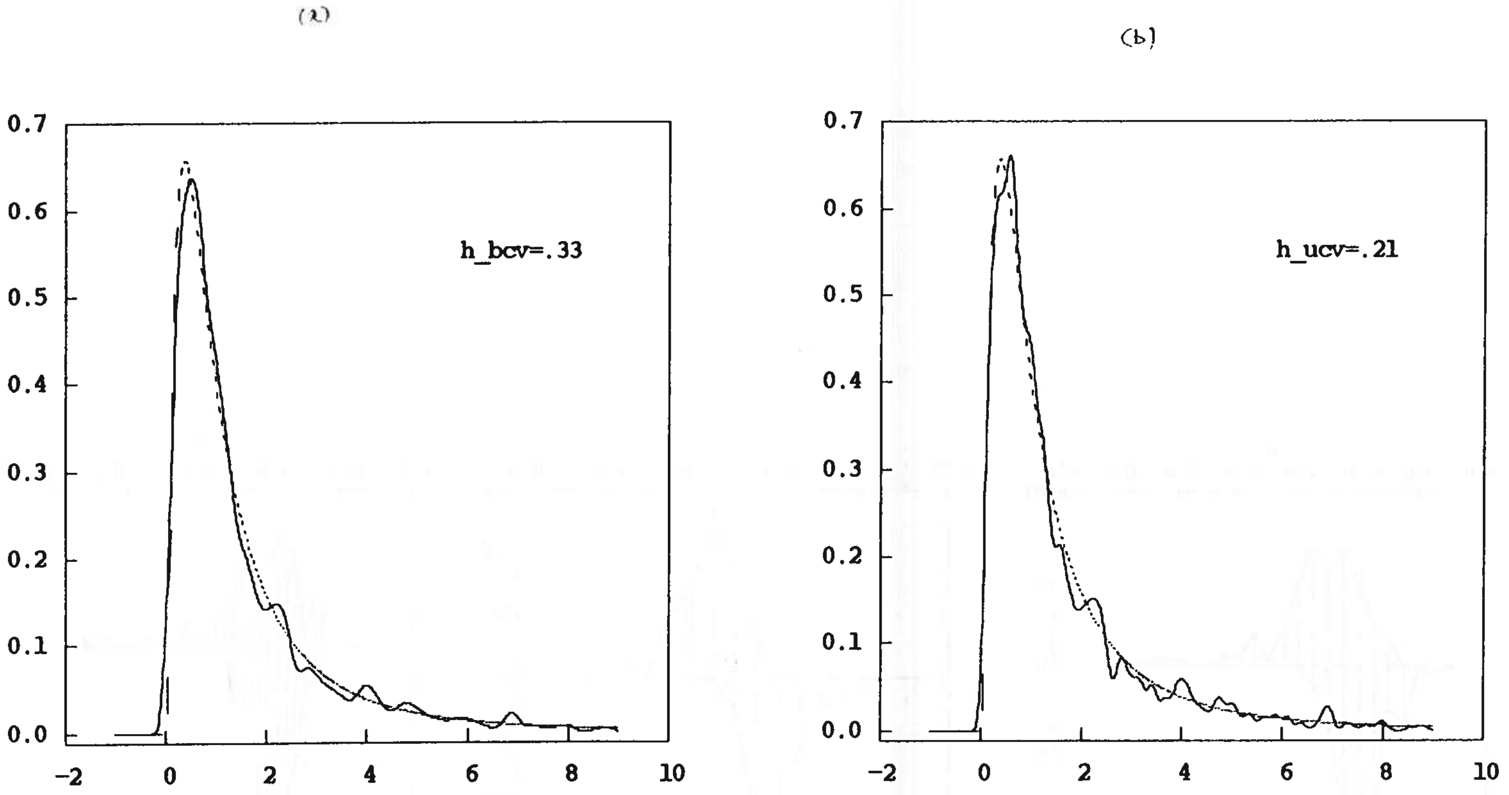
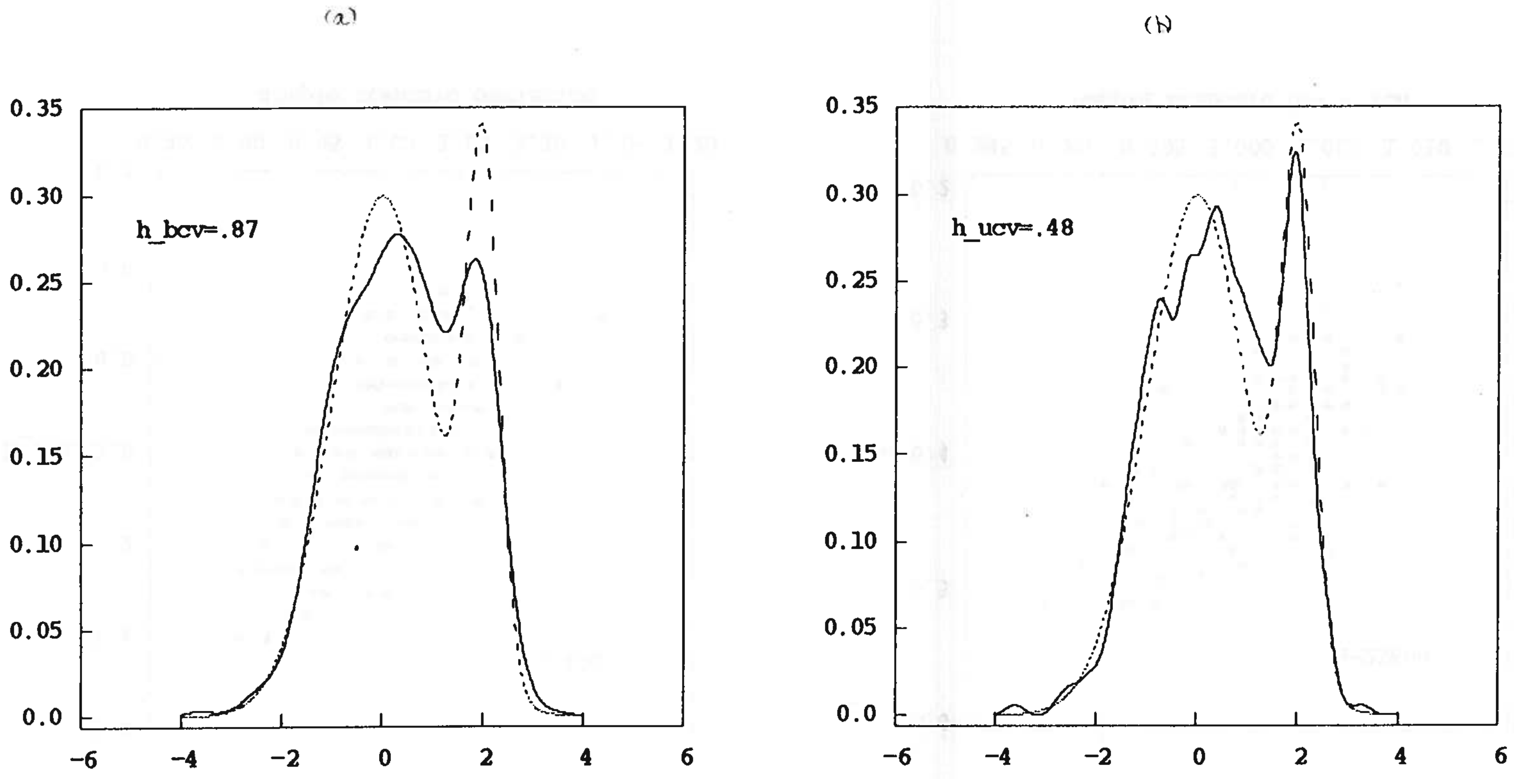


Figure 3



(a)

(b)

figure 7

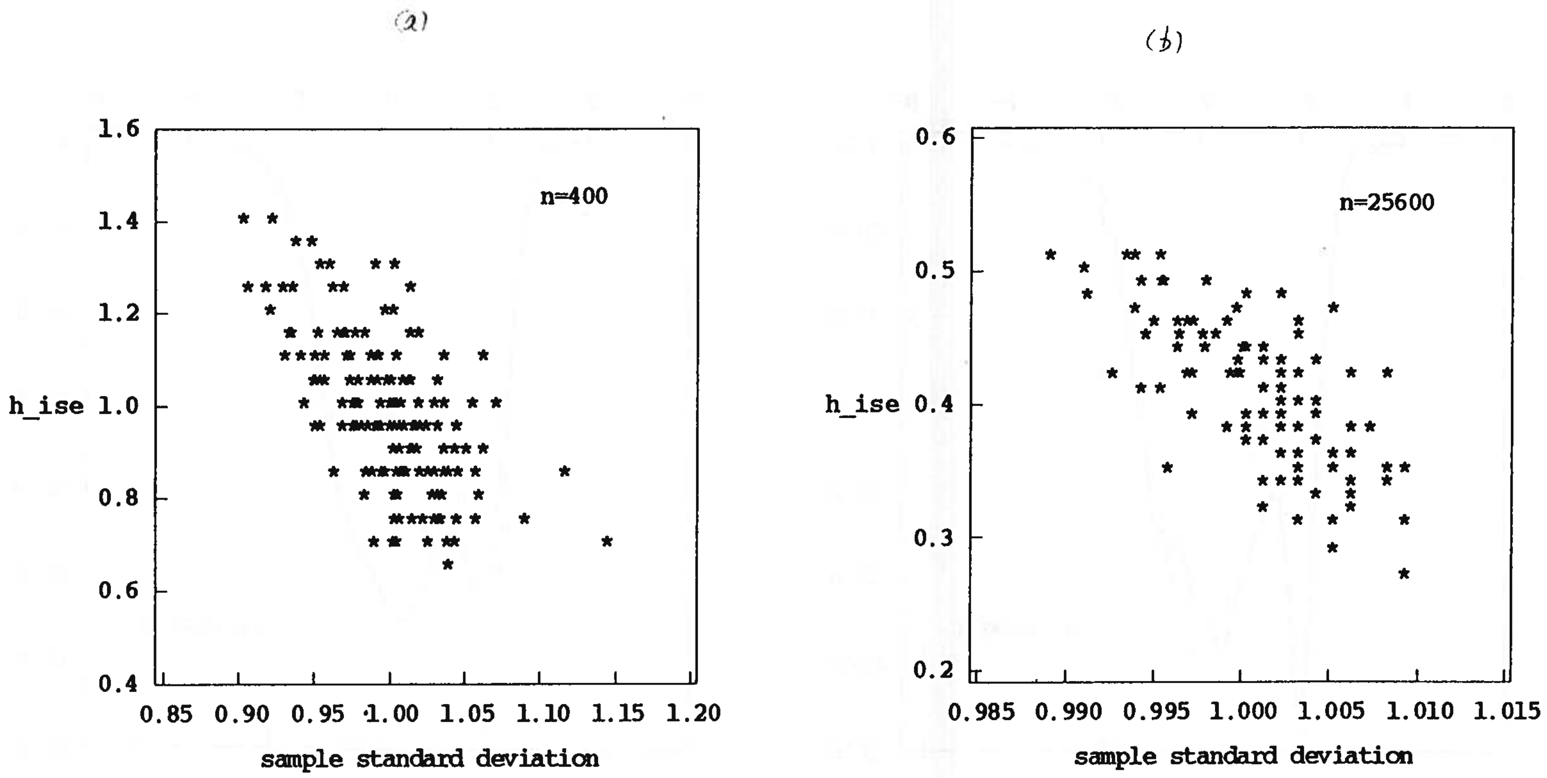
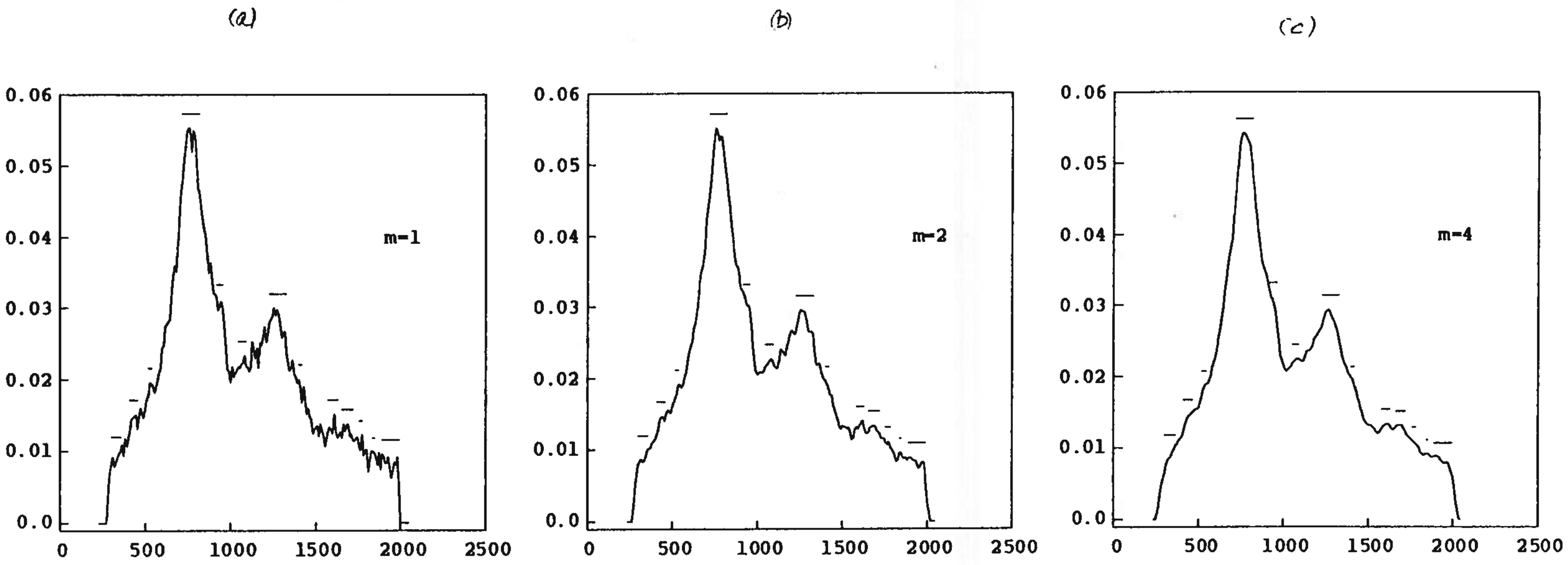

(a)
n=400
h_ise
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.85
0.90
0.95
1.00
1.05
1.10
1.15
1.20
sample standard deviation
(b)
n=25600
h_ise
0.6
0.5
0.4
0.3
0.2
0.985
0.990
0.995
1.000
1.005
1.010
1.015
sample standard deviation

Figure 13


(a)
(b)
(c)
0.06
0.05
0.04
0.03
0.02
0.01
0.0
0
500
1000
1500
2000
2500
m=1
m=2
m=4